The Dissertation Committee for James Robert Martin
certifies that this is the approved version of the following dissertation:

A Computational Framework for the Solution of Infinite-Dimensional Bayesian Statistical Inverse Problems with Application to Global Seismic Inversion

Committee:

Omar Ghattas, Supervisor

George Biros

Leszek Demkowicz

Sergey Fomel

Youssef Marzouk

Robert Moser

A Computational Framework for the Solution of Infinite-Dimensional Bayesian Statistical Inverse Problems with Application to Global Seismic Inversion

by

James Robert Martin, B.S., M.S.C.A.M.

DISSERTATION

Presented to the Faculty of the Graduate School of

The University of Texas at Austin

in Partial Fulfillment

of the Requirements

for the Degree of

DOCTOR OF PHILOSOPHY

THE UNIVERSITY OF TEXAS AT AUSTIN

August 2015

Dedicated to Kelly and Eliza

For our future

Acknowledgments

This dissertation would not have been possible without the collaboration, support, and inspiration from the many wonderful people in my life. These co-workers, friends, and family have not only helped me accomplish this body of work, but have also helped me become the person I am today. I can't possibly mention all of them, but I wish to acknowledge those who have significantly influenced my life.

I would like to acknowledge the many academic and professional colleagues who have contributed to my education, my research, and my life. All of these people are great friends who have provided critical contributions and perspective. I would not be who I am were it not for their influence.

- Omar Ghattas, my research supervisor - One of the most supportive and caring people I have ever known, he has treated me like family. I am constantly inspired by his compassion and enthusiasm and know that he will be a lifelong friend.

- Youssef Marzouk, my close friend and mentor - He helped me learn to navigate the world of uncertainty and opened doors to collaborate with researchers across the country. I have many fond memories of long evenings of detective work and exciting discoveries, and I look forward to many more.

- Georg Stadler, my close friend and collaborator - With patience, compassion and wisdom, he helped me return to work and renew faith in myself after I learned some of life's more painful lessons. He inspires me to be a stronger and kinder version of myself, and I look forward to many future discussions and collaborations.

- Noemi Petra, my close friend and collaborator - Her constant energy, cheerfulness, and willingness to help whenever possible have been a pleasure to work with. She worked tirelessly to produce a constant flow of new and interesting results, and being able to explore and consider these together made even the toughest days enjoyable.

- Lucas Wilcox, my close friend and collaborator - Beyond the excitement of exploring a field new to both of us, he provided immeasurable inspiration, support, and perseverance as I struggled to prepare my first major paper.

- Jerrold Marsden, my undergraduate thesis advisor at Caltech - He introduced me to the beauty and power of geometry, dynamical systems, and scientific computing.

- Paul Braterman, my TAMS chemistry professor - He took an active interest in my education, and enabled me to pursue scientific research early on.

- My co-authors Tan Bui, Carsten Burstedde, Tiangang Cui, Omar Ghattas, Youssef Marzouk, Noemi Petra, Antti Solonen, Alessio Spantini, Georg Stadler, Luis Tenorio, and Lucas Wilcox - I could not ask for a finer team of people to work with, and I am honored to count myself among them. Their dedication and work ethic is contagious, and I will never forget the shared triumphs, the countless hours of debugging, and the newfound understanding and perspective that we achieved together.

- Sue Rodriguez, friend and supporter - She went out of her way to help me whenever she could, and she frequently performed small logistical miracles. From insurance to travel to basic quality of life, her support enabled me to spend my energy where it was most important.

- Stephanie Rodriguez, CSEM Graduate Coordinator - For all of her help in my many transitions during my time in CSEM. She went above and beyond to minimize the stress on me.

- The wonderful people of CCGO, my academic family - A terrific group of talented and dedicated researchers. They both helped me and inspired me.

- The alumni, current fellows, and administrators of the DOE CSGF - I have had the wonderful privilege of joining this group of world class researchers in scientific computing. They gave me the freedom to pursue my passions in depth, and opened the doors for collaboration to fulfill them.

- My dissertation committee: George Biros, Leszek Demkowicz, Sergey Fomel, Omar Ghattas, Youssef Marzouk, and Robert Moser - They have provided invaluable perspective, feedback, and opportunities to present my work over the years. Their patience and accommodation throughout the course of my dissertation work has been greatly appreciated.

I would also like to acknowledge my family and friends. Their unconditional love and support carries me when I am weary, and helps me up when I fall down. Their constant faith in my worth and ability guides me to achieve my potential.

- Kelly Martin, my wife and companion - With her I have grown in ways I could never anticipate, and she often comprehends things about me before I do. Her love gives me life, her wisdom illuminates my world, and her strength keeps me going. She knows first-hand the worst and best of me, and still she pours her heart and soul into making us the best that we can be.

- Charlie Martin, the best brother a person could have - His generosity, love, and dedication have carried me through the toughest of times. Together we've built the routines and support structures on which I rely every day. I don't know what I would have done without him.

- Mary and Glen Martin, my parents - They worked hard to provide me the opportunities to pursue my passions, and I am eternally grateful for their encouragement, wisdom, guidance, love, and support. Mom, thank you for your enduring love, empathy, and comfort. So many of these qualities that I value deeply in myself, I learned from you. Dad, thank you for teaching me the joy of learning, the value of hard work, and the courage to forge ahead even when my path is uncertain.

- Joe Kirschvink, my uncle and mentor at Caltech - He offered me a home away from home, and helped me to expand my interests and nurture my continued growth as a scientist and engineer.

- Tony Toepfer, my closest friend - He is supportive, compassionate, caring, and "in it for the long haul." So few others understand the true experimental value of a well-placed oversized light-up penguin. He is always helping people, and I strive to emulate his readiness to improve the lives of the people around him.

- Nick Alger, my friend and colleague - Nick taught me the value of nature walks. We endlessly ponder new ways to think about mathematics, health, and the world around us, and then experiment with and engineer tools to make them better.

- Steve and Bing Brunton, and Paul Wali, my close friends from Caltech - Together, we learned that the possible is often easy, that the impossible is often only difficult, and that doing the impossible together creates lifelong friendships.

- Deborah Sharp, my confidant - She taught me how to combine the best of what is known with the best of what is uncertain, perhaps even unknowable, but still worth doing. She helped me learn how to live again.

To all of you and so many others, sincerely and wholeheartedly, thank you.

A Computational Framework for the Solution of Infinite-Dimensional Bayesian Statistical Inverse Problems with Application to Global Seismic Inversion

James Robert Martin, Ph.D.
The University of Texas at Austin, 2015

Supervisor: Omar Ghattas

Quantifying uncertainties in large-scale forward and inverse PDE simulations has emerged as a central challenge facing the field of computational science and engineering. The promise of modeling and simulation for prediction, design, and control cannot be fully realized unless uncertainties in models are rigorously quantified, since this uncertainty can potentially overwhelm the computed result. While statistical inverse problems can be solved today for smaller models with a handful of uncertain parameters, this task is computationally intractable using contemporary algorithms for complex systems characterized by large-scale simulations and high-dimensional parameter spaces. In this dissertation, I address issues regarding the theoretical formulation, numerical approximation, and algorithms for solution of infinite-dimensional Bayesian statistical inverse problems, and apply the entire framework to a problem in global seismic wave propagation.

Classical (deterministic) approaches to solving inverse problems attempt to recover the "best-fit" parameters that match given observation data, as measured in a particular metric. In the *statistical inverse problem*, we go one step further to return not only a point estimate of the best medium properties, but also a complete statistical description of the uncertain parameters. The result is a posterior *probability distribution* that describes our state of knowledge after learning from the available data, and provides a complete description of parameter uncertainty.

In this dissertation, a computational *framework* for such problems is described that wraps around the existing forward solvers, as long as they are appropriately equipped, for a given physical problem. Then a collection of tools, insights and numerical methods may be applied to solve the problem, and interrogate the resulting posterior distribution, which describes our final state of knowledge. We demonstrate the framework with numerical examples, including inference of a heterogeneous compressional wavespeed field for a problem in global seismic wave propagation with 10^6 parameters.

Table of Contents

Chapter 1

Framework Overview

In this chapter, I provide a high level overview of my dissertation research, and discuss the contributions of the remaining chapters toward the common goal of a computational framework for infinite-dimensional statistical inverse problems. Each subsequent chapter is adapted from an existing publication or a manuscript currently under review, and the majority of the technical discussion and literature review will be defered to the corresponding chapter. The relevant papers can be found here:

- A Stochastic Newton MCMC Method for Large-Scale Statistical Inverse Problems with Application to Seismic Inversion [143]
- A Computational Framework for Infinite-Dimensional Bayesian Inverse Problems Part I: The Linearized Case, with Application to Global Seismic Inversion [33]
- A Computational Framework for Infinite-Dimensional Bayesian Inverse Problems, Part II: Stochastic Newton MCMC with Application to Ice Sheet Flow Inverse Problems [166]
- Extreme-scale UQ for Bayesian inverse problems governed by PDEs [26]

- Optimal low-rank approximations of Bayesian linear inverse problems [186]
- Likelihood-informed dimension reduction for nonlinear inverse problems [57]

The following section describes an abstract infinite-dimensional inverse problem to set the stage for the abstract Bayesian statistical inverse problem which is described in section 1.3, and is of central interest to this work. Following this, I describe our contributions to this field, beginning with the framework for the linear Gaussian setting in section 1.4, and followed by the various extensions to the nonlinear setting in section 1.5. These nonlinear extensions are further separated into three basic approaches, each appropriate for solving a given class of statistical inverse problems: sampling with correction in section 1.5.1, implicit dimensionality reduction in section 1.5.2, and explicit dimensionality reduction in section 1.5.3.

1.1 The abstract deterministic inverse problem

Before we consider the statistical setting, it is informative to consider the deterministic inverse problem. The deterministic inverse problem has been well studied, and powerful insights and associated algorithms have enabled the solution of many large-scale deterministic inverse problems. In this section we briefly discuss these insights and the reasoning behind them, which we then apply to the statistical inverse problem in the following sections.

In the inverse problem setting, we seek to reconstruct an unknown heterogeneous *parameter field* over a physical problem domain $\Omega \subset \mathbb{R}^3$. We denote the parameter field by the function $m : \Omega \to \mathbb{R}$, so that for any physical point $\boldsymbol{x} \in \Omega$, $m(\boldsymbol{x})$ denotes the local value of the parameter field. The parameter m is taken to be an element of an infinite-dimensional function space. (The choice of function space and implications thereof are considered carefully in [33, 189]).

Generally, we cannot directly measure any property of the parameter m. Instead, we measure *observable data* $\boldsymbol{y}_{\text{obs}} \in \mathbb{R}^d$ which is indirectly related to the unknown parameter m. For simplicity, we assume the observable data to be finite-dimensional. A mathematical model of the process relating m to $\boldsymbol{y}_{\text{obs}}$ is encapsulated by the *parameter-to-observable map*

$$\boldsymbol{f}(m) \approx \boldsymbol{y}, \tag{1.1}$$

where $\boldsymbol{y} \in \mathbb{R}^d$ are the predicted observable quantities corresponding to the given parameter m. In practice, this relationship holds only approximately for several reasons—we may have model error related to the discrepancy between our mathematical model and the true underlying physics, we may have measurement error in the data $\boldsymbol{y}_{\text{obs}}$, and finally we may have numerical error in the computational simulation of $\boldsymbol{f}(m)$. For simplicity, we will combine all three sources of error into a single additive error term $\boldsymbol{e} \in \mathbb{R}^d$, so that

$$\boldsymbol{f}(m) = \boldsymbol{y} + \boldsymbol{e}. \tag{1.2}$$

The traditional deterministic setting formulates the inverse problem as a least squares optimization problem, in which the desired parameter is obtained as the solution m^* to an optimization problem where the observables most closely match the measured data, that is, to minimize the *data misfit* term:

$$m^* = \text{argmin}_m \frac{1}{2}\|\boldsymbol{f}(m) - \boldsymbol{y}_{\text{obs}}\|^2. \tag{1.3}$$

In general, the deterministic inverse problem is *ill-posed*, so that many parameters m will match the observed data equally (or almost equally) well. To uniquely select a single parameter, we must have some criterion to select the "best" one in some way. This metric for what is "best" is expressed in the form of a *regularization term* that is added to optimization problem we actually solve,

$$m^* = \text{argmin}_m \frac{1}{2}\|\boldsymbol{f}(m) - \boldsymbol{y}_{\text{obs}}\|^2 + \frac{1}{2}\|\mathcal{R}(m - m_0)\|^2, \tag{1.4}$$

so that we have a unique solution m^* with appropriate choice of the positive definite operator $\mathcal{R}$.

We next delve more deeply into the nature of the parameter-to-observable map $\boldsymbol{f}$. Generally, the parameter m is connected to the observations through the solution of a PDE parameterized by the unknown parameter m. For illustration, we consider that each m gives rise to a linear operator $\mathcal{K}(m)$, and we have

$$\mathcal{K}(m)u = q. \tag{1.5}$$

This is often referred to as the *forward problem*, mapping the parameter m to the *forward variable* u through solution of the given PDE. The right-hand side q is typically considered to be known, but dependence on the parameter m is possible if desired. The observable quantities $\boldsymbol{y}$ are connected to the solution of the PDE through the observation operator $\boldsymbol{\mathcal{B}}$,

$$\boldsymbol{y} = \boldsymbol{\mathcal{B}}u. \tag{1.6}$$

Finally, we may restate the deterministic inverse problem in the language of PDE-constrained optimization:

$$m^* = \text{argmin}_{m,u} \quad \frac{1}{2}\|\boldsymbol{\mathcal{B}}u - \boldsymbol{y}_{\text{obs}}\|^2 + \frac{1}{2}\|\mathcal{R}(m - m_0)\|^2 \tag{1.7a}$$

$$\text{subject to} \qquad \mathcal{K}(m)u = q. \tag{1.7b}$$

Solution to this deterministic inverse problem proceeds generally by the application of Newton's method, for which we desire to compute gradient and Hessian information for the given constrained optimization problem. We write the Lagrangian and introduce the Lagrange multiplier p, sometimes referred to as the *adjoint variable*:

$$\mathcal{L}(u, p, m) = \frac{1}{2}\|\boldsymbol{\mathcal{B}}u - \boldsymbol{y}_{\text{obs}}\|^2 + \frac{1}{2}\|\mathcal{R}(m - m_0)\|^2 + \langle p, \mathcal{K}(m)u - q\rangle. \tag{1.8}$$

The first order necessary terms for an optimal solution are thus:

$$0 = \delta_p\mathcal{L} \cdot \tilde{p} = \langle \tilde{p}, \mathcal{K}(m)u - q\rangle \qquad \forall\, \tilde{p}, \tag{1.9a}$$

$$0 = \delta_u\mathcal{L} \cdot \tilde{u} = \langle \tilde{u}, \boldsymbol{\mathcal{B}}^*(\boldsymbol{\mathcal{B}}u - \boldsymbol{y}_{\text{obs}})\rangle + \langle \tilde{u}, \mathcal{K}^*(m)p\rangle \qquad \forall\, \tilde{u}, \tag{1.9b}$$

$$0 = \langle g(m), \tilde{m}\rangle := \delta_m\mathcal{L} \cdot \tilde{m} = \langle p, [\delta_m\mathcal{K}(m) \cdot \tilde{m}]u\rangle + \langle \mathcal{R}\tilde{m}, \mathcal{R}(m - m_0)\rangle \qquad \forall\, \tilde{m}. \tag{1.9c}$$

Equation (1.9a) corresponds to the forward solve, equation (1.9b) corresponds to the so called adjoint solve, and finally we note that even at parameter values m away from the optimal solution, that is, when (1.9c) is not satisfied, the *reduced gradient* $g(m)$ is defined implicitly by this system of equations.

We next consider Newton's method to search for a parameter m^* which satisfies the system (1.9). To do this we need access to the derivative of the gradient, or the *Hessian* operator $H(m)$, and Newton's method proceeds iteratively using updates of the form:

$$m_{k+1} = m_k - [H(m)]^{-1} g(m). \tag{1.10}$$

Within each Newton iteration, we use an iterative solver (i.e., conjugate gradients) to compute the Newton step, and for this it is sufficient to be able to compute the action of the Hessian on any particular parameter direction $\hat{m}$, i.e., the *Hessian-vector product* $H(m)\hat{m} := \delta_m g(m) \cdot \hat{m}$. This can be derived by taking the variation of the system of equations (1.9) in the $\hat{m}$ direction:

$$0 = \langle \tilde{p}, [\delta_m \mathcal{K}(m) \cdot \hat{m}]u + \mathcal{K}(m)[\delta_m u \cdot \hat{m}] \rangle \qquad \forall\, \tilde{p}, \tag{1.11a}$$

$$0 = \langle \tilde{u}, \boldsymbol{\mathcal{B}}^* \boldsymbol{\mathcal{B}} [\delta_m u \cdot \hat{m}] \rangle + \langle \tilde{u}, [\delta_m \mathcal{K}^*(m) \cdot \hat{m}]p + \mathcal{K}^*(m)[\delta_m p \cdot \hat{m}] \rangle \quad \forall\, \tilde{u}, \tag{1.11b}$$

$$\begin{aligned} \langle H(m)\hat{m}, \tilde{m} \rangle =& \langle [\delta_m p \cdot \hat{m}], [\delta_m \mathcal{K}(m) \cdot \tilde{m}]u \rangle + \langle p, [\delta_m \mathcal{K}(m) \cdot \tilde{m}][\delta_m u \cdot \hat{m}] \rangle \\ &+ \langle p, [\delta^2_{mm} \mathcal{K}(m) \cdot (\tilde{m}, \hat{m})]u \rangle + \langle \mathcal{R}\tilde{m}, \mathcal{R}\hat{m} \rangle \qquad \forall\, \tilde{m}. \end{aligned} \tag{1.11c}$$

Next, we define the *incremental forward* $\hat{u}$ and *incremental adjoint* $\hat{p}$ variables as the variation of the corresponding forward and adjoint solutions,

$$\hat{u} := \delta_m u \cdot \hat{m} \qquad \text{and} \qquad \hat{p} := \delta_m p \cdot \hat{m}, \tag{1.12}$$

and observe that these can be found as the unknowns in the above system (1.11). Substituting in, we arrive at the standard expressions for the Hessian-vector product:

$$0 = \langle \tilde{p}, [\delta_m \mathcal{K}(m) \cdot \hat{m}]u + \mathcal{K}(m)\hat{u} \rangle \qquad \forall\, \tilde{p}, \tag{1.13a}$$

$$0 = \langle \tilde{u}, \boldsymbol{\mathcal{B}}^* \boldsymbol{\mathcal{B}} \hat{u} \rangle + \langle \tilde{u}, [\delta_m \mathcal{K}^*(m) \cdot \hat{m}]p + \mathcal{K}^*(m)\hat{p} \rangle \qquad \forall\, \tilde{u}, \tag{1.13b}$$

$$\begin{aligned} \langle H(m)\hat{m}, \tilde{m} \rangle = &\langle \hat{p}, [\delta_m \mathcal{K}(m) \cdot \tilde{m}]u \rangle + \langle p, [\delta_m \mathcal{K}(m) \cdot \tilde{m}]\hat{u} \rangle \\ &+ \langle p, [\delta^2_{mm} \mathcal{K}(m) \cdot (\tilde{m}, \hat{m})]u \rangle + \langle \mathcal{R}\tilde{m}, \mathcal{R}\hat{m} \rangle \qquad \forall\, \tilde{m}. \end{aligned} \tag{1.13c}$$

1.2 The discretized abstract inverse problem

Before proceeding with discussion of the statistical inverse problem setting, we discuss discretization of the problem for purposes of computation. The statistical problem can certainly be formulated in the infinite-dimensional setting as in the previous section, but for purposes of communication and intuition, we choose to present the statistical problem in the finite-dimensional setting. In doing so, some technical details will be omitted, but much of the intuition for this setting is essentially unchanged. For a more complete presentation of the infinite-dimensional setting, see [33, 189].

The parameter m is discretized using finite elements into a vector $\boldsymbol{m} \in \mathbb{R}^N$, where each element of the vector m_i represents the coefficient of one of the finite element basis functions $\phi_i(\boldsymbol{x})$. In addition, we choose the inner product for our finite-dimensional space to be the one induced from the $L^2(\Omega)$ function

space. Specifically, we define the *mass matrix* $\boldsymbol{M} \in \mathbb{R}^{N\times N}$ with components

$$M_{ij} := \int_\Omega \phi_i(\boldsymbol{x})\phi_j(\boldsymbol{x})d\boldsymbol{x}, \tag{1.14}$$

and the corresponding inner product is then given by

$$\langle \boldsymbol{m}_1, \boldsymbol{m}_2 \rangle := \boldsymbol{m}_1^T \boldsymbol{M} \boldsymbol{m}_2. \tag{1.15}$$

It is of course similarly necessary to discretize the solution variables u and p for computational purposes as well, but the framework does not depend on any particular numerical method for the forward simulation.

1.3 The abstract Bayesian statistical inverse problem

In this section, we describe the interpretation of the inverse problem in the Bayesian statistical framework, which is our ultimate problem of interest. In this setting, the uncertainty in the parameters recovered from the inverse problem is fully described by the *posterior probability distribution* with probability density function (pdf) $\pi_{\text{post}}(\boldsymbol{m} \mid \boldsymbol{y}_{\text{obs}})$, which ascribes to any parameter the relative probability that this might represent the "true" parameter (i.e., the parameter that actually produced the given observation data). The posterior pdf is obtained in Bayes' Theorem as the product of the pdfs for the *likelihood* and *prior* distributions, respectively.

The likelihood pdf $\pi_{\text{like}}(\boldsymbol{y}_{\text{obs}} \mid \boldsymbol{m})$ represents the probability that a given set of observations $\boldsymbol{y}_{\text{obs}}$ would be observed from a system with parameter $\boldsymbol{m}$. For this, we return to the consideration of the error in the parameter-to-observable map (1.2). A standard assumption which will be used throughout

this dissertation is that the error term $\boldsymbol{e}$ is a centered multivariate Gaussian with covariance matrix $\boldsymbol{\Gamma}_{\text{obs}}$. Then we may write $\boldsymbol{e} = \boldsymbol{f}(\boldsymbol{m}) - \boldsymbol{y} \sim \mathcal{N}(\boldsymbol{0}, \boldsymbol{\Gamma}_{\text{obs}})$, and the likelihood pdf is

$$\pi_{\text{like}}(\boldsymbol{y}_{\text{obs}} \mid \boldsymbol{m}) \propto \exp\left(-\frac{1}{2}\|\boldsymbol{f}(\boldsymbol{m}) - \boldsymbol{y}_{\text{obs}}\|^2_{\boldsymbol{\Gamma}^{-1}_{\text{obs}}}\right). \tag{1.16}$$

The normalization constant in the likelihood pdf, and by extension in the posterior pdf, is generally intractable to compute, and is not necessary for the computational framework described here. We observe before moving on that the negative log likelihood is similar in character to the data misfit term from the unregularized deterministic inverse problem seen in equation (1.3).

Next we consider the prior pdf, which describes our initial state of knowledge (or beliefs) about the parameter $\boldsymbol{m}$. We assume that the prior density is Gaussian, with mean $\boldsymbol{m}_0$ and covariance $\boldsymbol{\Gamma}_{\text{pr}}$,

$$\pi_{\text{pr}}(\boldsymbol{m}) \propto \exp\left(-\frac{1}{2}\|\boldsymbol{m} - \boldsymbol{m}_0\|^2_{\boldsymbol{\Gamma}^{-1}_{\text{pr}}}\right). \tag{1.17}$$

Again we note that the negative log prior is similar in character to the regularization term introduced in equation (1.4).

In a scientific setting, the assumption of Gaussianity of the prior distribution is a significant open question, and a proper treatment is well beyond the scope of this dissertation. We do note however that the restriction to Gaussian priors is less restrictive than it may initially seem, because the parameter-to-observable map $\boldsymbol{f}(\boldsymbol{m})$ can include arbitrary mappings from this parameter to a desired physically relevant quantity which itself may thus have a non-Gaussian

distribution. This is commonly utilized in problems where the parameter must be positive, and a log-normal prior is selected for the physical parameter.

Finally, Bayes' Rule provides the desired posterior pdf as

$$\pi_{\text{post}}(\boldsymbol{m} \mid \boldsymbol{y}_{\text{obs}}) \propto \pi_{\text{pr}}(\boldsymbol{m})\pi_{\text{like}}(\boldsymbol{y}_{\text{obs}} \mid \boldsymbol{m}), \tag{1.18}$$

and we therefore can state the Bayesian statistical inverse problem as the characterization of the posterior pdf

$$\pi_{\text{post}}(\boldsymbol{m} \mid \boldsymbol{y}_{\text{obs}}) \propto \exp\left(-\frac{1}{2}\|\boldsymbol{f}(\boldsymbol{m}) - \boldsymbol{y}_{\text{obs}}\|^2_{\boldsymbol{\Gamma}^{-1}_{\text{obs}}} - \frac{1}{2}\|\boldsymbol{m} - \boldsymbol{m}_0\|^2_{\boldsymbol{\Gamma}^{-1}_{\text{pr}}}\right). \tag{1.19}$$

The negative log posterior is therefore precisely the cost function from deterministic optimization (1.4), when the regularization is selected as the negative log prior and the data is weighted by the inverse noise covariance. This observation is the key insight that enables us to leverage knowledge and technology for the efficient solution of large-scale PDE-constrained optimization for use in the Bayesian statistical inverse problem. These tools and their adaptation to the Bayesian framework are discussed in the following section.

1.4 Framework for the linear Gaussian setting

We first consider the setting in which the parameter-to-observable map $\boldsymbol{f}$ is linear, or approximately linear over the support of the posterior distribution, so that we may write

$$\boldsymbol{f}(\boldsymbol{m}) = \boldsymbol{F}\boldsymbol{m}, \qquad \text{or} \qquad \boldsymbol{f}(\boldsymbol{m}) \approx \boldsymbol{f}(\boldsymbol{m}_{\text{MAP}}) + \boldsymbol{F}(\boldsymbol{m} - \boldsymbol{m}_{\text{MAP}}), \tag{1.20}$$

where $\boldsymbol{m}_{\text{MAP}}$ is the maximum a posteriori (MAP) estimate, the parameter that maximizes the posterior pdf

$$\boldsymbol{m}_{\text{MAP}} := \text{argmax}_{\boldsymbol{m}} \, \pi_{\text{post}}(\boldsymbol{m} \mid \boldsymbol{y}_{\text{obs}}), \tag{1.21}$$

which corresponds to the deterministic inverse solution $\boldsymbol{m}^*$ from equation (1.7). In the case where $\boldsymbol{f}$ is genuinely linear, this point $\boldsymbol{m}_{\text{MAP}}$ is also the exact mean of the posterior $\bar{\boldsymbol{m}}$. This completes the standard set of assumptions that characterize the *linear Gaussian setting*. To summarize, we have assumed the prior distribution to be Gaussian, the parameter-to-observable map to be linear over the support of the posterior distribution, and all sources of error in the parameter-to-observable map to be additive and Gaussian. Under these conditions, we can readily demonstrate the posterior distribution to be Gaussian by refactorization of the negative log posterior from equation

(1.19):

$$
\begin{aligned}
&-\log \pi_{\text{post}}(\boldsymbol{m} \mid \boldsymbol{y}_{\text{obs}})) && (1.22)\\
&\quad \approx \frac{1}{2}\|\boldsymbol{f}(\boldsymbol{m}_{\text{MAP}}) + \boldsymbol{F}(\boldsymbol{m} - \boldsymbol{m}_{\text{MAP}}) - \boldsymbol{y}_{\text{obs}}\|^2_{\boldsymbol{\Gamma}^{-1}_{\text{obs}}} + \frac{1}{2}\|\boldsymbol{m} - \boldsymbol{m}_0\|^2_{\boldsymbol{\Gamma}^{-1}_{\text{pr}}} + \text{const.} && (1.23)\\
&\quad = \frac{1}{2}\boldsymbol{m}^*\boldsymbol{F}\boldsymbol{\Gamma}^{-1}_{\text{obs}}\boldsymbol{F}\boldsymbol{m} + \boldsymbol{m}^*\boldsymbol{F}^*\boldsymbol{\Gamma}^{-1}_{\text{obs}}(\boldsymbol{f}(\boldsymbol{m}_{\text{MAP}}) - \boldsymbol{F}\boldsymbol{m}_{\text{MAP}} - \boldsymbol{y}_{\text{obs}}) + \text{const.}\\
&\qquad + \frac{1}{2}\boldsymbol{m}^*\boldsymbol{\Gamma}^{-1}_{\text{pr}}\boldsymbol{m} - \boldsymbol{m}^*\boldsymbol{\Gamma}^{-1}_{\text{pr}}\boldsymbol{m}_0 + \text{const.} && (1.24)\\
&\quad = \frac{1}{2}\boldsymbol{m}^*(\boldsymbol{F}\boldsymbol{\Gamma}^{-1}_{\text{obs}}\boldsymbol{F} + \boldsymbol{\Gamma}^{-1}_{\text{pr}})\boldsymbol{m}\\
&\qquad + \boldsymbol{m}^*[\boldsymbol{F}^*\boldsymbol{\Gamma}^{-1}_{\text{obs}}(\boldsymbol{f}(\boldsymbol{m}_{\text{MAP}}) - \boldsymbol{F}\boldsymbol{m}_{\text{MAP}} - \boldsymbol{y}_{\text{obs}}) - \boldsymbol{\Gamma}^{-1}_{\text{pr}}\boldsymbol{m}_0] + \text{const.} && (1.25)\\
&\quad = \frac{1}{2}\boldsymbol{m}^*\boldsymbol{\Gamma}^{-1}_{\text{post}}\boldsymbol{m} - \boldsymbol{m}^*\boldsymbol{\Gamma}^{-1}_{\text{post}}\bar{\boldsymbol{m}} + \text{const.} && (1.26)\\
&\quad = \frac{1}{2}\|\boldsymbol{m} - \bar{\boldsymbol{m}}\|^2_{\boldsymbol{\Gamma}^{-1}_{\text{post}}} + \text{const.} && (1.27)
\end{aligned}
$$

Our computational framework is therefore tasked with characterization of the Gaussian posterior distribution:

$$
\begin{aligned}
&\pi_{\text{post}}(\boldsymbol{m} \mid \boldsymbol{y}_{\text{obs}})) \propto \exp\left(-\frac{1}{2}\|\boldsymbol{m} - \bar{\boldsymbol{m}}\|^2_{\boldsymbol{\Gamma}^{-1}_{\text{post}}}\right), && (1.28a)\\
&\text{where} \quad \boldsymbol{\Gamma}_{\text{post}} = (\boldsymbol{F}^*\boldsymbol{\Gamma}_{\text{obs}}\boldsymbol{F} + \boldsymbol{\Gamma}^{-1}_{\text{pr}})^{-1} && (1.28b)\\
&\text{and} \quad \bar{\boldsymbol{m}} = \boldsymbol{\Gamma}_{\text{post}}[\boldsymbol{F}^*\boldsymbol{\Gamma}^{-1}_{\text{obs}}(\boldsymbol{f}(\boldsymbol{m}_{\text{MAP}}) - \boldsymbol{F}\boldsymbol{m}_{\text{MAP}} - \boldsymbol{y}_{\text{obs}}) - \boldsymbol{\Gamma}^{-1}_{\text{pr}}\boldsymbol{m}_0]. && (1.28c)
\end{aligned}
$$

Even with these explicit expressions for the posterior pdf, significant challenges remain. Below, we describe these challenges and the insights utilized in our computational framework to overcome them. We present this computational framework in detail for the linear setting in [33] for a global seismic inversion problem of parameter dimension $N \approx 10^5$, and demonstrate scala-

bility of the framework in [26] to a similar problem of parameter dimension $N \approx 10^6$.

Proper specification of the infinite-dimensional inverse problem is nontrivial. Our computational framework follows the approach of [189] to ensure our infinite-dimensional probability spaces are well-defined and satisfy some basic sanity criteria (e.g., we expect bounded variance for pointwise parameter values). To this end, we restrict our attention to priors constructed using the elliptic differential operator of the form

$$\mathcal{A}m := -\alpha\nabla \cdot (\boldsymbol{\Theta}\nabla m) + \alpha m, \tag{1.29}$$

and after appropriate choice of boundary conditions, the prior covariance operator is specified as $\mathcal{C} = \mathcal{A}^{-\gamma}$. This allows us to specify the amount of smoothness in samples from the prior distribution (by selecting the order of the differential operator γ), connects with theory that guarantees the resulting covariance operator is trace-class [189], and finally provides a connection with the Matérn covariance functions frequently used in geostatistics [137], including the ability to freely specify non-stationarity (via inhomogeneity in α and $\boldsymbol{\Theta}$) and anisotropy (via anisotropy in $\boldsymbol{\Theta}$). For the large-scale computations in [26, 33], we further restrict the prior to the setting where $\gamma = 2$, which allows for direct access to the square root of the prior covariance, effected by a single elliptic solve.

Proper discretization is essential. While this at some level is just a matter of not making mistakes in the derivation, we have found such mistakes

very easy to make, and correspondingly difficult to detect. For example, the correct adjoint for an operator involving the parameter space is not simply its matrix transpose; the adjoint must take into account the mass-weighted inner product defined in equation (1.15). One nonintuitive consequence of this is that the covariance matrix of an i.i.d. vector of standard normal random variables is $\boldsymbol{M}$ rather than the usual identity $\boldsymbol{I}$, and indeed the procedure for prior and posterior sample generation makes use of the operator $\boldsymbol{M}^{-1/2}$.

Our framework presents a pattern for discretization such that the expressions for each of the finite-dimensional quantities directly parallels its analog in the infinite-dimensional setting. This facilitates the derivation of additional desired quantities of interest (we make no claim to compute all of them), and helps provide concrete intuition for the infinite-dimensional quantities that may be unfamiliar. A detailed presentation of this is provided in [33] and reviewed briefly again in [166].

Efficient manipulation of the prior covariance operator is essential. Because we have specified this using the inverse of an elliptic differential operator, storage and application of this operator is accomplished using sparse operators and modern $O(N)$ solvers for elliptic problems (e.g., algebraic multigrid). Additionally, direct specification of the square root operator as opposed to the full operator enables multiple computations in the computational framework that are otherwise difficult and potentially expensive computationally. For small and medium scale statistical inverse problems

(e.g., those in [57, 143, 166]), direct factorization of the prior covariance is reasonable, but without use of the scalable methods described here, the computations in [26, 33, 84] would not have been tractable.

Efficient manipulation of the parameter-to-observable map is essential. By far the most significant computational expense in large-scale statistical inverse problems occurs in the evaluation of the forward problem and its derivatives. We therefore seek to evaluate these only when necessary, and to extract maximum benefit from them when we do. To this end, we leverage techniques from state of the art algorithms in large-scale (deterministic) optimization, including inexact Newton optimization, conjugate gradient (CG) solvers for the Newton step, backtracking Armijo linesearch, and adjoint PDE solves (i.e., those in (1.9b) and (1.13b)). These techniques have been demonstrated for many problems to converge in a number of iterations independent of the underlying mesh (i.e., the parameter dimension N).

The key insight that enables this mesh independent convergence is that the data misfit Hessian,

$$\boldsymbol{H}_{\text{misfit}} := \boldsymbol{F}^* \boldsymbol{\Gamma}_{\text{obs}}^{-1} \boldsymbol{F}, \tag{1.30}$$

is often a compact operator. Iterative Krylov subspace solvers such as CG will then interact only with the finite-dimensional range space of the Hessian, which contains precisely those parameter modes informed by the data. This compactness property of the Hessian is due to the fact that the observation data can only inform a limited number of modes in parameter space. The rank

of the linearized $\boldsymbol{F}$ is bounded by the dimension d of the data space for finite-dimensional data, and is usually much smaller for realistic statistical inverse problems. This happens for a number of reasons: practical constraints restrict where and what kind of measurements can be made, instrument sensitivity and environmental noise restricts the confidence with which we can measure certain quantities (e.g., high frequency perturbations in time or space), and physical characteristics of the forward problem can dilute or obscure the information content of most observations (e.g. for wave problems: damping of high frequency oscillations, shadow zones behind strong reflectors where waves do not reach, and simultaneous arrivals of many waves from different sources).

To extend this insight to the statistical inverse problem setting, we consider the *prior-preconditioned data misfit Hessian*,

$$\tilde{\boldsymbol{H}} := \boldsymbol{\Gamma}_{\text{pr}}^{1/2} \boldsymbol{F}^* \boldsymbol{\Gamma}_{\text{obs}}^{-1} \boldsymbol{F} \boldsymbol{\Gamma}_{\text{pr}}^{1/2}, \tag{1.31}$$

for which the range space describes precisely the set of parameter modes which are informed by the data, but also not already "known" by the prior distribution. If we are already very certain in our beliefs (expressed by the prior), it takes a correspondingly large amount of evidence (expressed by the data) to significantly alter or further confirm those original beliefs. The prior-preconditioned data misfit Hessian serves to balance the contribution from these two effects, and its range space describes precisely the modes in parameter space that we expect to change in the posterior distribution with respect to the prior.

Because $\boldsymbol{H}_{\text{misfit}}$ is compact and $\boldsymbol{\Gamma}_{\text{pr}}$ is taken as the inverse of an elliptic differential operator, $\tilde{\boldsymbol{H}}$ is also compact and often admits a low-rank decomposition. Our framework thus proceeds by factoring the Hessian (i.e., the inverse posterior covariance from equation (1.28b)) to expose $\tilde{\boldsymbol{H}}$,

$$\begin{aligned}
\boldsymbol{H} &= \boldsymbol{\Gamma}_{\text{post}}^{-1} && (1.32)\\
&= \boldsymbol{F}^{*}\boldsymbol{\Gamma}_{\text{obs}}^{-1}\boldsymbol{F} + \boldsymbol{\Gamma}_{\text{pr}}^{-1} && (1.33)\\
&= \boldsymbol{\Gamma}_{\text{pr}}^{-1/2}\Big[\boldsymbol{\Gamma}_{\text{pr}}^{1/2}\boldsymbol{F}^{*}\boldsymbol{\Gamma}_{\text{obs}}^{-1}\boldsymbol{F}\boldsymbol{\Gamma}_{\text{pr}}^{1/2} + \boldsymbol{I}\Big]\boldsymbol{\Gamma}_{\text{pr}}^{-1/2} && (1.34)\\
&= \boldsymbol{\Gamma}_{\text{pr}}^{-1/2}\Big[\tilde{\boldsymbol{H}} + \boldsymbol{I}\Big]\boldsymbol{\Gamma}_{\text{pr}}^{-1/2}, && (1.35)
\end{aligned}$$

and then replacing $\tilde{\boldsymbol{H}}$ with its low-rank approximation computed by Lanczos or randomized SVD. We can see from the above factorization that only eigenvalues that are significant compared to 1 are required for accuracy, and we need only compute until the desired level of tolerance is reached (typical values for us are eigenvalue thresholds of 10^{-1} or 10^{-2}).

We first described this factorization for the finite-dimensional setting in [143], extended it to the infinite-dimensional setting in [33], and the low-rank decompsition of $\tilde{\boldsymbol{H}}$ is a primary focus of the theoretical results in [186].

Optimal approximation of the posterior distribution is essential. We have gone to great lengths to ensure that we avoid all unnecessary computation, but we must also be certain that we are not omitting computation that is essential. (Or, in the setting where we might be forced to undersolve, that our framework returns the most significant results before the less

significant ones.) By utilizing the eigendecomposition of (1.31), we are able to order the eigenfunctions of parameter space by the extent to which they are affected by the update from prior to posterior, with the eigenfunctions corresponding to the largest eigenvalues most affected. It is then reasonable to expect that for a fixed rank r, the best approximation of (1.31) yields the optimal approximation of the exact posterior after the update is performed. In [186], this is proven to be the case for the linear setting as measured in distribution by both Hellinger distance and Kullback-Leibler divergence.

Post-processing of results is nontrivial. Finally, after the low-rank representation of the posterior distribution is obtained, post-processing to obtain the desired output visualizations or quantities of interest may still be nontrivial. For example, generating samples from the posterior distribution usually requires a form of square root factorization for the posterior covariance, such as $\boldsymbol{\Gamma}_{\text{post}} = \boldsymbol{L}\boldsymbol{L}^*$. Similarly, visualization of a pointwise variance field essentially requires the extraction of the diagonal of $\boldsymbol{\Gamma}_{\text{post}}$ in a particular basis.

While our framework cannot hope to anticipate all possible computations that a user might desire, many can be performed directly from the low-rank representation of the posterior distribution. In [33], we provide explicit expressions for the efficient computation of the above two quantities for demonstration, and of course Monte Carlo estimation using samples is always a possibility for difficult quantities of interest.

1.5 Framework for the non-Gaussian setting

In this section, we provide an overview for the work which addresses the more general setting in which the parameter-to-observable map may be nonlinear, so that in general the resulting posterior distribution is not Gaussian. We present three classes of algorithms designed to build on the efficient computational techniques and optimality results surveyed in the previous section, and describe the conditions under which each will be most applicable.

First, if the posterior distribution is not Gaussian, but still well approximated by the Gaussian approximation constructed at the MAP estimate, we can sample from this approximate distribution and apply a correction based on the evaluation of the true posterior at each sample point.

For more difficult problems, we describe two complementary approaches, which we describe as implicit and explicit dimensionality reduction, respectively. In the implicit schemes, we consider Markov-chain Monte Carlo methods tailored to the underlying low-dimensional structure of the problem, and demonstrate that the resulting algorithms are able to sample from large-scale non-Gaussian distributions while retaining the full parameter dimension. Alternatively, an explicit dimensionality reduction approach first computes a global low-rank basis, in which the action of the parameter-to-observable map is well approximated. Statistical inversion can then proceed using the coefficients of this reduced basis as the new auxiliary parameters, and finally the full posterior is reconstructed using Rao-Blackwellization.

1.5.1 Sampling with correction

For some problems, it is observed that the posterior distribution is quite well approximated by the linearized Gaussian distribution. It is perhaps surprising that this would ever be the case for a realistic statistical inverse problem with a complicated parameter-to-observable map, but it is not actually unreasonable to anticipate in some settings. In particular, consider problems in which the data noise in equation (1.2) is very small, or similarly by the law of large numbers, many independent observations inform the same underlying parameters. In this case, even if the parameter-to-observable map is globally nonlinear, it may still be nearly linear as a mapping onto the subset of the observable space on which the likelihood has non-negligible probability density, resulting in a nearly Gaussian posterior distribution.

In [84], we consider two methods by which we can utilize samples from the linearized Gaussian posterior to sample the true posterior, namely independence sampler Markov-chain Monte Carlo, and importance sampling. Both approaches require that we evaluate the pdf for the true posterior at each sample point, but no further intrusive (i.e., gradient or Hessian) information is required, and the required computations are embarassingly parallel.

First, we can use our approximate posterior as a *proposal distribution* for the Metropolis-Hastings (MH) algorithm 1, which results in an independence sampler Markov-chain Monte Carlo (MCMC) method. In each iteration of the MH algorithm, the current state of the Markov chain is $\boldsymbol{m}_k$, and a sample is generated from the proposal density $q(\boldsymbol{m}_k, \boldsymbol{m}_{\text{proposal}})$. This proposal

Algorithm 1 Metropolis-Hastings Algorithm to sample density $\pi(\boldsymbol{m}_k)$

Choose initial $\boldsymbol{m}_1$, set $\boldsymbol{m}_k = \boldsymbol{m}_1$
Compute $\pi(\boldsymbol{m}_k)$
for $k = 2, ..., N$ **do**
  Draw sample $\boldsymbol{m}_{\text{proposal}}$ from the proposal density $q(\boldsymbol{m}_k, \boldsymbol{m}_{\text{proposal}})$
  Compute $\pi(\boldsymbol{m}_{\text{proposal}})$
  Compute $\alpha(\boldsymbol{m}_k, \boldsymbol{m}_{\text{proposal}}) = \min\left(1, \frac{\pi(\boldsymbol{m}_{\text{proposal}})q(\boldsymbol{m}_{\text{proposal}}, \boldsymbol{m}_k)}{\pi(\boldsymbol{m}_k)q(\boldsymbol{m}_k, \boldsymbol{m}_{\text{proposal}})}\right)$
  Draw $u \sim \mathcal{U}([0, 1])$
  if $u < \alpha(\boldsymbol{m}_k, \boldsymbol{m}_{\text{proposal}})$ **then**
    Accept: Set $\boldsymbol{m}_k = \boldsymbol{m}_{\text{proposal}}$
  else
    Reject: do nothing ($\boldsymbol{m}_k = \boldsymbol{m}_{k-1}$)
  end if
end for

$\boldsymbol{m}_{\text{proposal}}$ is subsequently subjected to an accept/reject step that ensures the resulting sample chain converges to the correct posterior distribution. Here, we simply choose the approximate posterior distribution (1.28) for the MH proposal, independent of the current state $\boldsymbol{m}_k$ of the Markov chain. As with all MCMC methods, the challenges of convergence assessment and burn-in removal still remain and are problem dependent, and must be addressed as such. We note that the performance of the MCMC method can be assessed in part by the acceptance probability $\alpha(\boldsymbol{m}_k, \boldsymbol{m}_{\text{proposal}})$. If it were precisely the case that $\pi(\boldsymbol{m}_{\text{proposal}}) = q(\boldsymbol{m}_k, \boldsymbol{m}_{\text{proposal}})$, then we would have $\alpha(\boldsymbol{m}_k, \boldsymbol{m}_{\text{proposal}}) = 1$, every sample would be accepted, and we would have perfect MCMC convergence. This is the reason we expect the indepedence sampler to converge quickly when the approximate posterior is very near to the true posterior.

Alternatively, samples from the approximate posterior distribution can

be understood in the context of importance sampling. Here, we keep all of the original samples generated from the approximate distribution, but assign to each a sample weight based on the ratio between the approximate posterior and true posterior pdf values. To this end, consider the following estimator for the expected value of a given quantity of interest, $z(\boldsymbol{m})$:

$$\mathbb{E}[z(\boldsymbol{m})] = \int z(\boldsymbol{m})\ \pi(\boldsymbol{m}) d\boldsymbol{m} \tag{1.36}$$

$$= \int z(\boldsymbol{m}) \frac{\pi(\boldsymbol{m})}{\tilde{\pi}(\boldsymbol{m})}\ \tilde{\pi}(\boldsymbol{m}) d\boldsymbol{m} \tag{1.37}$$

$$\approx \frac{1}{N\overline{w}} \sum_{k=1}^{N} z(\boldsymbol{m}_k) w(\boldsymbol{m}_k), \tag{1.38}$$

$$\text{where} \qquad w(\boldsymbol{m}_k) = \frac{\pi(\boldsymbol{m}_k)}{\tilde{\pi}(\boldsymbol{m}_k)} \qquad \text{and} \qquad \overline{w} = \frac{1}{N} \sum_{k=1}^{N} w(\boldsymbol{m}_k), \tag{1.39}$$

and where the samples $\{\boldsymbol{m}_k\}_{k=1,\ldots,N}$ are generated from the approximate posterior distribution with density $\tilde{\pi}(\boldsymbol{m})$. Thus, by interpreting the quantity $w(\boldsymbol{m}_k)$ as a weight for each sample $\boldsymbol{m}_k$, we can treat these as samples from the desired posterior distribution, and Monte Carlo estimates for the desired quantities of interest can be computed using (1.38), provided that the surrogate quantity of interest $z(\boldsymbol{m}_k)\frac{\pi(\boldsymbol{m}_k)}{\tilde{\pi}(\boldsymbol{m}_k)}$ has bounded variance under the approximate posterior distribution.

1.5.2 Implicit Dimensionality Reduction: MCMC approaches

When the linearization at the MAP estimate does not provide a good approximation of the parameter-to-observable map, independence sampler MCMC and importance sampling approaches may converge too slowly to sufficiently

characterize the true posterior distribution, given a practical limit on available computing resources. In this setting, we seek to design MCMC algorithms tailored to the local structure of the posterior distribution, using local derivative information to guide the sampling process.

In [143], we describe the stochastic Newton MCMC method, which is tailored to the underlying local structure of the posterior distribution by constructing the Gaussian matching the local gradient and curvature information at the current MCMC point. Specifically, the proposal used in the Metropolis-Hastings algorithm 1 is computed by

$$\boldsymbol{m}_{\text{proposal}} = \boldsymbol{m}_k - \boldsymbol{H}(\boldsymbol{m}_k)^{-1}\boldsymbol{g}(\boldsymbol{m}_k) + \boldsymbol{H}(\boldsymbol{m}_k)^{-1/2}\boldsymbol{n}, \tag{1.40}$$

where $\boldsymbol{g}(\boldsymbol{m}_k)$ and $\boldsymbol{H}(\boldsymbol{m}_k)$ are the gradient and Hessian of the negative log posterior as computed in (1.9) and (1.13), and $\boldsymbol{n}$ is a sample from a standard normal distribution in $\mathbb{R}^N$. We can interpret this as the stochastic version of the Newton step in equation (1.10), hence the name *stochastic Newton*.

In general this method is very expensive – every proposed sample requires a full Hessian evaluation in every MCMC iteration, even if the proposed sample is subsequently rejected. Still, in performance comparisons [143] with a state of the art blackbox algorithm (DRAM) and a preconditioned (using the prior covariance) Langevin MCMC method (which makes use of gradient information), stochastic Newton was reasonably well converged within twelve hours of computation, whereas neither DRAM nor Langevin were able to adequately converge to the correct distribution even after weeks of computation.

We argue that this is due to the fact that the additional curvature information provided by the Hessian is crucial for reasonable performance in high dimensions.

The primary obstacles for stochastic Newton are its high computational cost, and potential for loss of robustness for very nonlinear problems. When the local curvature changes significantly between the current and proposed sample points, the back transition probability $q(\boldsymbol{m}_{\text{proposal}}, \boldsymbol{m}_k)$ in the MH algorithm 1 is often negligible, leading to poor performance of stochastic Newton MCMC. In [166], we consider a variant of stochastic Newton that attempts to overcome these obstacles. In this method, we evaluate the Hessian a single time at the MAP estimate, and reuse the MAP Hessian for all subsequent MCMC samples. This reduces the cost of each MCMC sample to a single evaluation of the posterior density and its gradient at the proposed sample point, and also proves to be more robust for many problems, due to increased predictability for the behavior of the back transition probability in MH. For the experiments in [166], this method was the clear winner in terms of both MCMC and overall computational performance.

1.5.3 Explicit Dimensionality Reduction: Global Reduced Basis approaches

Finally, we consider explicit dimensionality reduction of the large-scale statistical inverse problem. Here, our focus is slightly different than in the approaches outlined in the previous sections. We do not necessarily seek to solve

the statistical inverse problem as such; rather we seek to determine a global reduced basis in which most or all of the update from prior to posterior takes place. Once such a basis is identified, the number of effective parameters for the statistical inverse problem will be much smaller, enabling the use of many established algorithms suitable for low-dimensional problems. Our objective here is simply to enable the use of these algorithms for the large-scale problems of interest in this dissertation.

In [57], we outline an approach which builds on the optimality of the reduced subspace identified by the prior-preconditioned misfit Hessian described in [186], and which seeks to identify an analog of this subspace that applies globally to the posterior distribution instead of as the result of a local linearization at a point. In this method, we propose a bootstrap algorithm to generate samples from the posterior based on our current version of the global reduced subspace, compute the local reduced subspace at this new sample point, and subsequently fold the new information into our approximation of the global reduced subspace. The process is terminated after enough iterations proceed without significant change to the global reduced subspace.

Finally, after characterization of the posterior distribution within this reduced subspace is complete, the full space may be reconstructed using Rao-Blackwellization, which effectively utilizes the assumption that the posterior distribution has not changed from the prior distribution in directions orthogonal to the global reduced basis. This results in a representation of the posterior that is conditionally Gaussian in the orthogonal directions, and may facilitate

certain post-processing computations. As one example, in [57], we consider estimates obtained via Monte Carlo averaging, and argue that the variance in these estimates may be greatly reduced by sampling only within the reduced subspace, and performing analytic integration in the orthogonal directions.

Chapter 2

A computational framework for infinite-dimensional Bayesian inverse problems. Part I: The linearized case, with application to global seismic inversion

The content of this chapter is based on an existing publication[1] which is joint work with Tan Bui-Thanh, Georg Stadler, and my advisor Omar Ghattas. Georg contributed most of the effort in setting up the deterministic wave propagation and observation operators for the numerical experiments in this chapter. Georg and I collaborated on the implementation of the algorithms for statistical inversion, and finally interpretation and visualization of the results. All authors had significant contribution to the remaining content of this chapter.

Abstract

We present a computational framework for estimating the uncertainty in the numerical solution of linearized infinite-dimensional statistical inverse

[1] T. Bui-Thanh, O. Ghattas, J. Martin, and G. Stadler. A computational framework for infinite-dimensional Bayesian inverse problems Part I: The linearized case, with application to global seismic inversion. *SIAM Journal on Scientific Computing*, 35(6):A2494–A2523, 2013. http://epubs.siam.org/doi/abs/10.1137/12089586X

problems. We adopt the Bayesian inference formulation: given observational data and their uncertainty, the governing forward problem and its uncertainty, and a prior probability distribution describing uncertainty in the parameter field, find the posterior probability distribution over the parameter field. The prior must be chosen appropriately in order to guarantee well-posedness of the infinite-dimensional inverse problem and facilitate computation of the posterior. Furthermore, straightforward discretizations may not lead to convergent approximations of the infinite-dimensional problem. And finally, solution of the discretized inverse problem via explicit construction of the covariance matrix is prohibitive due to the need to solve the forward problem as many times as there are parameters.

Our computational framework builds on the infinite-dimensional formulation proposed by Stuart [189], and incorporates a number of components aimed at ensuring a convergent discretization of the underlying infinite-dimensional inverse problem. The framework additionally incorporates algorithms for manipulating the prior, constructing a low rank approximation of the data-informed component of the posterior covariance operator, and exploring the posterior that together ensure scalability of the entire framework to very high parameter dimensions. We demonstrate this computational framework on the Bayesian solution of an inverse problem in 3D global seismic wave propagation with hundreds of thousands of parameters.

2.1 Introduction

We present a scalable computational framework for the quantification of uncertainty in large scale *inverse problems*; that is, we seek to estimate probability densities for uncertain parameters,[2] given noisy observations or measurements and a model that maps parameters to output observables. The forward problem—which, without loss of generality, we take to be governed by PDEs—is usually well-posed (the solution exists, is unique, and is stable to perturbations in inputs), causal (later-time solutions depend only on earlier time solutions), and local (the forward operator includes derivatives that couple nearby solutions in space and time). The inverse problem, on the other hand, reverses this relationship by seeking to estimate uncertain parameters from measurements or observations. The great challenge of solving inverse problems lies in the fact that they are usually ill-posed, non-causal, and non-local: many different sets of parameter values may be consistent with the data, and the inverse operator couples solution values across space and time.

Non-uniqueness stems in part from the sparsity of data and the uncertainty in both measurements and the PDE model itself, and in part from non-convexity of the parameter-to-observable map. The popular approach to obtaining a unique "solution" to the inverse problem is to formulate it as an optimization problem: minimize the misfit between observed and predicted

[2] We use the term *parameters* broadly to describe general model inputs that may be subject to uncertainty, which might include model parameters, boundary conditions, initial conditions, sources, geometry, and so on.

outputs in an appropriate norm while also minimizing a *regularization* term that penalizes unwanted features of the parameters. Estimation of parameters using the regularization approach to inverse problems as described above will yield an estimate of the "best" parameter values that simultaneously fit the data and minimize the regularization penalty term. However, we are interested not just in point estimates of the best-fit parameters, but a *complete statistical description* of the parameters values that is consistent with the data. The *Bayesian* approach [117, 192] does this by reformulating the inverse problem as a problem in *statistical inference*, incorporating uncertainties in the observations, the parameter-to-observable map, and prior information on the parameters. The solution of this inverse problem is the *posterior* probability distribution of the parameters, which reflects the degree of confidence in their values. Thus we are able to quantify the resulting uncertainty in the parameters, taking into account uncertainties in the data, model, and prior information.

The inverse problems we target here are characterized by *infinite dimensional* parameter fields. This presents multiple difficulties, including proper choice of prior to guarantee well-posedness of the infinite-dimensional inverse problem, proper discretization to assure convergence to solutions of the infinite-dimensional problem, and algorithms for constructing and manipulating the posterior covariance matrix that insure scalability to very large parameter dimensions. The approach we adopt in this paper follows [189], which seeks to first fully specify the statistical inverse problem on the infinite-dimensional

parameter space. In order to accomplish this goal, we postulate the prior distribution as a Gaussian random field with covariance operator given by the square of the inverse of an elliptic PDE. This choice ensures that samples of the parameter field are (almost surely) continuous as functions, and that the statistical inverse problem is well-posed. To achieve a finite-dimensional approximation to the infinite-dimensional solution, we carefully construct a function-space-aware discretization of the parameter space.

The remaining challenge presented by infinite-dimensional statistical inverse problems is in computing statistics of the (discretized) posterior distribution. This is notoriously challenging for inverse problems governed by expensive-to-solve forward problems and high-dimensional parameter spaces (as in our application to global seismic wave propagation in Section 2.6). The difficulty stems from the fact that evaluation of the probability of each point in parameter space requires solution of the forward PDE problem (which can take many hours on a large supercomputer), and many such evaluations (millions or more) are required to adequately sample the (discretized) posterior density in high dimensions by conventional Markov-chain Monte Carlo (MCMC) methods. In complementary work [143], we are developing methods that accelerate MCMC sampling of the posterior by employing a local Gaussian approximation of the posterior as a proposal density, which is computed from the Hessian of the negative log posterior. Here, as an alternative, we consider the case of the linearized inverse problem; by linearization we mean that the parameter-to-observable map is linearized about the point that maximizes the posterior,

which is known as the maximum *a posteriori* (MAP) point. With this linearization, the posterior becomes Gaussian, and its mean is given by the MAP point; this can be found by solving an appropriately weighted regularized nonlinear least squares optimization problem. Furthermore, the posterior covariance matrix is given by the Hessian of the negative log posterior evaluated at the MAP point.

Unfortunately, straightforward computation of the—nominally dense—Hessian is prohibitive, requiring as many forward-like solves as there are uncertain parameters (which in our example problem in Section 2.6, is hundreds of thousands). However, the data are typically informative about a low dimensional subspace of the parameter field: that is, the Hessian of the data misfit term is a compact operator that is sparse with respect to some basis. We exploit this fact to construct a low rank approximation of the (prior preconditioned) data misfit Hessian using matrix-free Lanczos iterations [73,143], which we observe to require a dimension-independent number of iterations. Each iteration requires a Hessian-vector product, which amounts to just a pair of forward/adjoint PDE solves, as well as a prior covariance operator application. Since we take the prior covariance in the form of the inverse of an elliptic differential operator, its application can be computed scalably via multigrid. The Sherman-Morrison-Woodbury formula is then invoked to express the covariance of the posterior. Finally, we show that the resulting expressions necessary for visualization and interrogation of the posterior distribution require just elliptic PDE solves and vector sums and inner products.

In particular, the corresponding dense operators are never formed or stored. Solving the statistical inverse problem thus reduces to solving a fixed number of forward and adjoint PDE problems as well as an elliptic PDE representing the action of the prior. Thus, when the forward PDE problem can be solved in a scalable manner (as it is for our seismic wave propagation example in Section 2.6), the entire computational framework is scalable with respect to forward problem dimension, uncertain parameter field dimension, and data dimension.

The computational framework presented here is applied to a sequence of realistic large-scale 3D Bayesian inverse problems in global seismology, in which the acoustic wavespeed of an unknown heterogeneous medium is to be inferred from noisy waveforms recorded at sparsely located receivers. Numerical results are presented for several problems with the number of unknown parameters up to 431,000. We have employed a similar approach for problems with more than one million parameters in related work [26].

In the following sections, we provide an overview of the framework for infinite-dimensional Bayesian inverse problems following [189] (Section 2.2), present a consistent discretization scheme (Section 2.3) for the infinite-dimensional problem, summarize a method for computing the MAP point (Section 2.4), describe our low rank-based covariance approximation (Section 2.5), and present results of the application of our framework to the Bayesian solution of an inverse problem in 3D global seismic wave propagation (Section 2.6).

2.2 Bayesian framework for infinite-dimensional inverse problems

2.2.1 Overview

In the Bayesian formulation, we state the inverse problem as a problem of *statistical inference* over the space of parameters. The solution of the resulting statistical inverse problem is a posterior probability distribution that reflects our degree of confidence that any set of candidate parameters might contain the actual values that gave rise to the data via the model and were consistent with the prior information. Bayes' formula, presented in its infinite dimensional form in Section 2.2.2, defines this posterior probability distribution by combining a prior probability distribution with a likelihood model.

The inversion parameter is a function assumed to be defined over an open, bounded, and sufficiently regular set $\Omega \subset \mathbb{R}^3$. The statistical inverse problem is therefore naturally posed in an appropriate function space setting. Here, we adopt the infinite-dimensional framework developed in [189]. In particular, we choose a prior that ensures sufficient regularity of the parameter as required for the statistical inverse problem to be well-posed. We will represent the prior as a Gaussian random field whose covariance operator is the inverse of an elliptic differential operator. For certain problems, non-Gaussian priors can be important, but the use of non-Gaussian priors in statistical inverse problems is still subject to active research, in particular for infinite-dimensional parameters. Thus, here we restrict ourselves to priors given by Gaussian random fields. Let us motivate the choice of the covariance operator as inverse

of an elliptic differential operator by considering two alternatives. A common choice for covariance operators in statistical inverse problems with a moderate number of parameters is to specify the covariance function, which gives the covariance of the parameter field between any two points. This necessitates either construction and "inversion" of a dense covariance matrix or expansion in a truncated Karhunen-Loéve (KL) basis. In the large-scale setting, inversion of a dense covariance matrix is clearly intractable, and the truncated KL approach can be impractical since it may require many terms to prevent biasing of the solution toward the strong prior modes. On the contrary, specifying the covariance as the inverse of an elliptic differential operator enables us to build on existing fast solvers for elliptic operators without constructing the dense operator. Discretizations of elliptic operators often satisfy a conditional independence property, which relates them to Gaussian Markov random fields and allows for statistical interpretation [16, 182]. Even if a Gaussian Markov random field is not based on an elliptic differential operator, this Markov property permits the use of fast, sparsity-exploiting algorithms for instance for taking samples from the distribution, [181]. Our implementation employs multigrid as solver for the discretized elliptic systems.

A useful prior distribution must have bounded variance and have meaningful realizations. In our infinite-dimensional setting, we require samples to be pointwise well-defined, for instance, continuous. Furthermore, it is convenient to have the ability to apply the square root of the covariance operator, e.g., this is used to compute samples from a Gaussian distribution. We con-

sider a Gaussian random field m on a domain $\Omega \subset \mathbb{R}^3$ with mean m_0 and covariance function $c(\boldsymbol{x}, \boldsymbol{y})$ describing the covariance between $m(\boldsymbol{x})$ and $m(\boldsymbol{y})$

$$c(\boldsymbol{x}, \boldsymbol{y}) = \mathbb{E}\left[(m(\boldsymbol{x}) - m_0(\boldsymbol{x}))(m(\boldsymbol{y}) - m_0(\boldsymbol{y}))\right] \quad \text{for } \boldsymbol{x}, \boldsymbol{y} \in \Omega. \tag{2.1}$$

The corresponding covariance operator $\mathcal{C}_0$ is

$$\left(\mathcal{C}_0 \phi\right)(\boldsymbol{x}) = \int_\Omega c\left(\boldsymbol{x}, \boldsymbol{y}\right) \phi\left(\boldsymbol{y}\right) \, d\boldsymbol{y} \tag{2.2}$$

for sufficiently regular functions ϕ defined over Ω. Thus, if the covariance operator is given by the solution operator of an elliptic PDE, the covariance function is the corresponding Green's function. Thus, Green's function properties have direct implications for properties of the random field m. For instance, since Green's functions of the Laplacian in one spatial dimension are bounded, the random field with the Laplacian as covariance operator is of bounded variance. However, in two and three space dimensions, Green's functions $c(\boldsymbol{x}, \boldsymbol{y})$ of the Laplacian are singular along the diagonal, and thus the corresponding distribution has unbounded variance. Thus, intuitively the PDE solution operator used as covariance operator $\mathcal{C}_0$ has to be sufficiently smoothing and have bounded Green's functions. Indeed, this is necessary for the well-posedness of the infinite-dimensional Bayesian formulation [189]. The biharmonic operator, for example, has bounded Green's functions in two and three space dimensions. We choose $\mathcal{C}_0 = \mathcal{A}^{-2}$, where $\mathcal{A}$ is a Laplacian-like operator specified in Section 2.2.3. This provides the desired simple and fast-to-apply square root operator $\mathcal{C}_0^{1/2} = \mathcal{A}^{-1}$ and allows a straightforward discretization.

An approach to extract information from the posterior distribution is to find the maximum a posterior (MAP) point, which amounts to the solution of an optimization problem as summarized in Section 2.2.4. Finally, in Section 2.2.5, we introduce a linearization of the parameter-to-observable map. This results in a Gaussian approximation of the posterior, which is the main focus of this paper.

2.2.2 Bayes' formula in infinite dimensions

To define Bayes' formula, we require a likelihood function that defines, for a given parameter field m, the distribution of observations $\boldsymbol{y}^{\text{obs}}$. Here, we assume a finite-dimensional vector $\boldsymbol{y}^{\text{obs}} \in \mathbb{R}^q$ of such observations. We introduce the *parameter-to-observable map* $\boldsymbol{f} : X := L^2(\Omega) \to \mathbb{R}^q$ as a deterministic function mapping a parameter field m to so-called observables $\boldsymbol{y} \in \mathbb{R}^q$, which are predictions of the observations. For the problems targeted here, an evaluation of $\boldsymbol{f}(m)$ requires a PDE solve followed by the application of an observation operator to extract $\boldsymbol{y}$ from the PDE solution. Even when the parameter m coincides with the "true" parameter, the observables $\boldsymbol{y}$ may still differ from the measurements $\boldsymbol{y}^{\text{obs}}$ due to measurement noise and inadequacy (i.e., the lack of fidelity of the governing PDEs with respect to reality) of the parameter-to-observable map $\boldsymbol{f}$. As is common practice, we assume the discrepancy between $\boldsymbol{y}$ and $\boldsymbol{y}^{\text{obs}}$ to be described by a Gaussian additive noise $\boldsymbol{\eta} \sim \mu_{\text{noise}} = \mathcal{N}(0, \boldsymbol{\Gamma}_{\text{noise}})$, independent of m. In particular, we have

$$\boldsymbol{y}^{\text{obs}} = \boldsymbol{f}(m) + \boldsymbol{\eta}, \tag{2.3}$$

which allows us to write the likelihood probability density function (pdf) as

$$\pi_{\text{like}}(\boldsymbol{y}^{\text{obs}}|m) \propto \exp\left(-\frac{1}{2}(\boldsymbol{f}(m) - \boldsymbol{y}^{\text{obs}})^T \boldsymbol{\Gamma}_{\text{noise}}^{-1}(\boldsymbol{f}(m) - \boldsymbol{y}^{\text{obs}})\right). \tag{2.4}$$

The Bayesian solution to the infinite-dimensional inverse problem is then defined as follows: given the likelihood π_{like} and the prior measure μ_0, find the conditional measure μ^y of m that satisfies the Bayes' formula

$$\frac{d\mu^y}{d\mu_0} = \frac{1}{Z}\pi_{\text{like}}(\boldsymbol{y}^{\text{obs}}|m), \tag{2.5}$$

where $Z = \int_X \pi_{\text{like}}(\boldsymbol{y}^{\text{obs}}|m)\, d\mu_0$ is a normalization constant. The formula (2.5) is understood as the Radon-Nikodym derivative of the posterior probability measure μ^y with respect to the prior measure μ_0. In order for (2.5) to be well defined, $\boldsymbol{f} : X \to \mathbb{R}^q$ is assumed to be locally Lipschitz and quadratically bounded in the sense of Assumption 2.7 in [189]. While the Bayes' formula (2.5) is valid in finite and infinite dimensions, a more intuitive form of Bayes' formula that uses Lebesgue measures and thus only holds in finite dimensions is given in Section 2.3.5.

2.2.3 Parameter space and the prior

As discussed in the introduction of Section 2.2, we use a squared inverse elliptic operator as covariance operator $\mathcal{C}_0$ in (2.1), i.e., $\mathcal{C}_0 = \mathcal{A}^{-2}$. We first specify the elliptic PDE corresponding to $\mathcal{A}$ in weak form. For $s \in L^2(\Omega)$, the solution $m = \mathcal{A}^{-1}s$ satisfies

$$\alpha \int_\Omega (\boldsymbol{\Theta}\nabla m) \cdot \nabla p + mp\, d\boldsymbol{x} = \int_\Omega sp\, d\boldsymbol{x} \text{ for all } p \in H^1(\Omega), \tag{2.6}$$

with $\alpha > 0$, and $\boldsymbol{\Theta}(\boldsymbol{x}) \in \mathbb{R}^{3\times 3}$ is symmetric, uniformly bounded, and positive definite. Note that for $s \in L^2(\Omega)$, there exists a unique solution $m \in H^1(\Omega)$ by the Lax-Milgram theorem. Since $s \in L^2(\Omega)$ in (2.6), regularity results, e.g. [9, 68], show that in fact $m \in H^2(\Omega)$ provided $\partial\Omega$ is sufficiently smooth, e.g., Ω is a $C^{1,1}$ domain. In this case, (m, s) satisfies the elliptic differential equation

$$-\alpha\nabla\cdot(\boldsymbol{\Theta}\nabla m) + \alpha m = s \quad \text{in } \Omega, \tag{2.7a}$$

$$\alpha(\boldsymbol{\Theta}\nabla m)\cdot\boldsymbol{n} = 0 \quad \text{on } \partial\Omega, \tag{2.7b}$$

where $\boldsymbol{n}$ denotes the outward unit normal on $\partial\Omega$.

Let us denote by $\mathcal{A}$ the differential operator together with its domain of definition specified by (2.7); hence $\mathcal{A}$ is a densely defined operator on $L^2(\Omega)$ with the following domain

$$D(\mathcal{A}) := \left\{ m \in H^2(\Omega) : \alpha\boldsymbol{\Theta}\nabla m\cdot\mathbf{n} = 0 \right\}.$$

The operator $\mathcal{A}$ is assumed to be "Laplacian-like" in the sense of Assumption 2.9 in [189]. In brief, this assumption requires that $\mathcal{A}$ be positive definite, self-adjoint, invertible, and have eigenfunctions that form an orthonormal basis of $L^2(\Omega)$. Additionally, certain growth conditions on the eigenvalues and $L^\infty(\Omega)$ norms of the eigenfunctions are enforced[3].

[3]We note that this growth condition on the eigenfunctions may not be straightforward to demonstrate (or may not even hold) for a non-rectangular domain Ω and nonconstant coefficient $\boldsymbol{\Theta}$. In these cases, we expect that alternative proofs of the results in [189] can be accessed via regularity properties of the covariance function for the prior distribution. See for example [2, 137].

To summarize, we consider m as a Gaussian random field whose distribution law is a Gaussian measure $\mu_0{:=}\mathcal{N}(m_0, \mathcal{C}_0)$ on $\boldsymbol{L}^2(\Omega)$, with mean $m_0 \in D(\mathcal{A})$ and covariance operator $\mathcal{C}_0{:=}\mathcal{A}^{-2}$. The definition of the Gaussian prior measure is meaningful since $\mathcal{A}^{-2}$ is a trace class operator on $L^2(\Omega)$ [189], which guarantees bounded variance and almost surely pointwise well-defined samples since $\mu_0(X) = 1$ holds, where $X := C(\Omega)$ denotes the space of continuous functions defined on Ω (see [189, Lemma 6.25]).

2.2.4 The MAP point

As a first step in exploring the solution of the statistical inverse problem, we determine the maximum a posteriori (MAP) estimate of the posterior measure. In a finite-dimensional setting, the MAP estimate is the point in parameter space that maximizes the posterior probability density function. This notion does not generalize directly to the infinite-dimensional setting, but we can still define the MAP estimate m_{MAP} as the point m in parameter space that asymptotically maximizes the measure of a ball with radius ε centered at m, in the limit as $\varepsilon \to 0$. We recall that the Cameron-Martin space E equipped with the inner product $(\cdot\,,\cdot)_E := (\mathcal{C}_0^{-1/2}\cdot, \mathcal{C}_0^{-1/2}\cdot)$ associated with $\mathcal{C}_0$ is the range of $\mathcal{C}_0^{1/2}$ [97], and hence coincides with $D(\mathcal{A})$. Using variational arguments, it can be shown (see [189]) that m_{MAP} is given by solving the optimization problem

$$\min_{m \in E} \mathcal{J}(m)\,, \tag{2.8}$$

where

$$\mathcal{J}(m) := \frac{1}{2}\left\|\boldsymbol{f}(m) - \boldsymbol{y}^{\text{obs}}\right\|^2_{\boldsymbol{\Gamma}^{-1}_{\text{noise}}} + \frac{1}{2}\left\|\mathcal{A}(m - m_0)\right\|^2_{L^2(\Omega)}. \tag{2.9}$$

The well-posedness of the optimization problem (2.8) is guaranteed by the assumptions on $\boldsymbol{f}(m)$ in Section 2.2.2.

2.2.5 A linearized Bayesian formulation

Once we have obtained the MAP estimate m_{MAP}, we approximate the parameter-to-observable map $\boldsymbol{f}(m)$ by its linearization about m_{MAP}, which ultimately results in a Gaussian approximation to the posterior distribution, as shown below. When the parameter-to-observable map is nearly linear this is a reasonable approximation; moreover, there are other scenarios in which the linearization, and the resulting Gaussian approximation, may be useful. Of particular interest here are the limits of small data noise and many observations. In the small noise case, the parameter-to-obvervable map can be nearly linear as a mapping into the subset of the observable space on which the likelihood distribution is non-negligible—even when $\boldsymbol{f}(m)$ is significantly nonlinear. The asymptotic normality discussions in [80, 127] suggest that under certain conditions, the many observations case can lead to a Gaussian posterior. Finally, even if this approximation fails to describe the posterior distribution adequately, the linearization is still useful in building an initial step for the rejection sampling approach or a Gaussian proposal distribution for the Metropolis-Hastings algorithm [143, 174]. These methods are related to the sampling algorithm in [86], which also employs derivative information

to respect the local structure of the parameter space.

Assuming that the parameter-to-observable map $\boldsymbol{f}$ is Fréchet differentiable, we linearize the right hand side of 2.3 around m_{MAP} to obtain

$$\boldsymbol{y}^{\text{obs}} \approx \boldsymbol{f}\left(m_{\text{MAP}}\right) + \boldsymbol{F}\left(m - m_{\text{MAP}}\right) + \boldsymbol{\eta}$$

where $\boldsymbol{F}$ is the Fréchet derivative of $\boldsymbol{f}\left(m\right)$ evaluated at m_{MAP}. Consequently, the posterior distribution μ^{y} of m conditional on $\boldsymbol{y}^{\text{obs}}$ is a Gaussian measure $\mathcal{N}\left(m_{\text{MAP}}, \mathcal{C}_{\text{post}}\right)$ with mean m_{MAP} and covariance operator $\mathcal{C}_{\text{post}}$ defined by [189]:

$$\mathcal{C}_{\text{post}} = \left(\boldsymbol{F}^{\natural}\boldsymbol{\Gamma}_{\text{noise}}^{-1}\boldsymbol{F} + \mathcal{C}_{0}^{-1}\right)^{-1}, \tag{2.10}$$

with $\boldsymbol{F}^{\natural}$ denoting the adjoint of $\boldsymbol{F}$, an operator from the space of observations $\mathbb{R}^{q}$ to $L^{2}\left(\Omega\right)$. In principle, a local Gaussian approximation of the posterior at the MAP point can also be found for non-Gaussian priors and when the noise in the observables is not additive and Gaussian as in (2.3). In these cases, however, even for a linear parameter-to-observable map the local Gaussian approximation might only be reasonable approximation to the true posterior distribution in a small neighborhood around the MAP point.

2.3 Discretization of the Bayesian inverse problem

2.3.1 Overview

Next, we present a numerical discretization of the infinite-dimensional Bayesian statistical inverse problem described in Section 2.2.2. The discretized

parameter space is inherently high-dimensional (with dimension dependent upon the mesh size). If discretization is not performed carefully at each step, it is unlikely that the discrete solutions will converge to the desired infinite-dimensional solution in a meaningful way [125, 189].

In the following, and particularly in Section 2.3.3, we choose a mass matrix-weighted vector product instead of the standard Euclidean vector product. While this is a natural choice in finite element discretizations [19,199], this does lead to a few complications, for instance, the use of covariance operators that are not symmetric in the conventional sense (they are self-adjoint however). This choice is much better suited for proper discretization of the infinite-dimensional expressions given in this paper, and the resulting numerical expressions for computation will more closely resemble their infinite-dimensional counterparts in Section 2.2. By contrast, the correct corresponding expressions in the Euclidean inner product are significantly less intuitive in our opinion, and ultimately more cumbersome to manipulate and interpret than the development we give here.

We provide finite-dimensional approximations of the prior and the posterior distributions in Sections 2.3.4 and 2.3.5, respectively. To study and visualize the uncertainty in Gaussian random fields, such as the prior and posterior distributions, we generate realizations (i.e., samples) and compute pointwise variance fields. This must be done carefully in light of the mass-weighted inner products due to the finite element discretization introduced in Section 2.3.3. We present explicit expressions for computing these quantities

for the prior in the Sections 2.3.6 and 2.3.7. The fast generation of samples and the pointwise variance field from the Gaussian approximation of the posterior exploits the low rank ideas presented in Section 2.5. Thus, the presentation of the corresponding expressions is postponed to Section 2.5.3.

2.3.2 Finite-dimensional parameter space

We consider a finite-dimensional subspace V_h of $L^2(\Omega)$ originating from a finite element discretization with continuous Lagrange basis functions $\{\phi_j\}_{j=1}^n$, which correspond to the nodal points $\{\boldsymbol{x}_j\}_{j=1}^n$, such that

$$\phi_j(\boldsymbol{x}_i) = \delta_{ij}, \qquad \text{for } i,j \in \{1,\dots,n\}.$$

Instead of statistically inferring parameter functions $m \in L^2(\Omega)$, we perform this task on the approximation $m_h = \sum_{j=1}^n m_j \phi_j \in V_h$. Consequently, the coefficients $(m_1,\dots,m_n)^T \in \mathbb{R}^n$ are the actual parameters to be inferred. For simplicity of notation, we shall use the boldface symbol $\boldsymbol{m} = (m_1,\dots,m_n)^T$ to denote the nodal vector of a function m_h in V_h.

2.3.3 Discrete inner product

Since we postulate the prior Gaussian measure on $L^2(\Omega)$, the finite-dimensional space V_h inherits the L^2-inner product. Thus, inner products between nodal coefficient vectors must be weighted by a mass matrix $\boldsymbol{M}$ to approximate the infinite-dimensional L^2-inner product. We denote this weighted inner product by $(\cdot\,,\cdot)_{\boldsymbol{M}}$ and assume that $\boldsymbol{M} \in \mathbb{R}^{n\times n}$ is symmetric and positive definite. To distinguish $\mathbb{R}^n$ with the $\boldsymbol{M}$-weighted inner product from the

usual Euclidean space $\mathbb{R}^n$, we denote it by $\mathbb{R}^n_{\boldsymbol{M}}$. For any $m_1, m_2 \in L^2(\Omega)$, observe that $(m_1, m_2)_{L^2(\Omega)} \approx (m_{1h}, m_{2h})_{L^2(\Omega)} = (\boldsymbol{m}_1, \boldsymbol{m}_2)_{\boldsymbol{M}} = \boldsymbol{m}_1^T \boldsymbol{M} \boldsymbol{m}_2$, which motivates the choice of $\boldsymbol{M}$ as the finite element mass matrix defined by

$$M_{ij} = \int_\Omega \phi_i(\boldsymbol{x}) \phi_j(\boldsymbol{x}) d\boldsymbol{x} \ , \quad i, j \in \{1, \ldots, n\} \, . \tag{2.11}$$

When using the $\boldsymbol{M}$-inner product, there is a critical distinction that must be made between the matrix adjoint and the matrix transpose. For an operator $\boldsymbol{B} : \mathbb{R}^n_{\boldsymbol{M}} \to \mathbb{R}^n_{\boldsymbol{M}}$, we denote the matrix transpose by $\boldsymbol{B}^T$ with entries $[B^T]_{ij} = B_{ji}$. The adjoint $\boldsymbol{B}^*$ of $\boldsymbol{B}$, however, must satisfy

$$(\boldsymbol{B}^* \boldsymbol{m}_1, \boldsymbol{m}_2)_{\boldsymbol{M}} = (\boldsymbol{m}_1, \boldsymbol{B} \boldsymbol{m}_2)_{\boldsymbol{M}} \quad \text{for all } \boldsymbol{m}_1, \boldsymbol{m}_2 \in \mathbb{R}^n_{\boldsymbol{M}}. \tag{2.12}$$

This implies that

$$\boldsymbol{B}^* = \boldsymbol{M}^{-1} \boldsymbol{B}^T \boldsymbol{M}. \tag{2.13}$$

In the following, we also need the adjoints $\boldsymbol{F}^\natural$ of $\boldsymbol{F} : \mathbb{R}^n_{\boldsymbol{M}} \to \mathbb{R}^q$ and $\boldsymbol{V}^\diamond$ of $\boldsymbol{V} : \mathbb{R}^r \to \mathbb{R}^n_{\boldsymbol{M}}$ (for some r), where $\mathbb{R}^q$ and $\mathbb{R}^r$ are endowed with the Euclidean inner product. The desired adjoints can be can be expressed as

$$\boldsymbol{F}^\natural = \boldsymbol{M}^{-1} \boldsymbol{F}^T, \tag{2.14}$$

$$\boldsymbol{V}^\diamond = \boldsymbol{V}^T \boldsymbol{M}. \tag{2.15}$$

Next, let P_h be the projection from $L^2(\Omega)$ to V_h. Then, the matrix representation $\boldsymbol{B} : \mathbb{R}^n_{\boldsymbol{M}} \to \mathbb{R}^n_{\boldsymbol{M}}$ for the operator $\mathcal{B}_h := P_h \mathcal{B} P_h'$, where $\mathcal{B} : L^2(\Omega) \to L^2(\Omega)$ and $P_h' : V_h \to L^2(\Omega)$, is implicitly given with respect to the Lagrange basis $\{\phi_i\}_{i=1}^n$ in V_h by

$$\int_\Omega \phi_i \mathcal{B} \phi_j \, dx = (\boldsymbol{e}_i, \boldsymbol{B} \boldsymbol{e}_j)_{\boldsymbol{M}},$$

where $\boldsymbol{e}_i$ is the coordinate vector corresponding to the basis function ϕ_i. As a result, one can write $\boldsymbol{B}$ explicitly as

$$\boldsymbol{B} = \boldsymbol{M}^{-1}\boldsymbol{K}, \tag{2.16}$$

where $\boldsymbol{K}$ is given by

$$K_{ij} = \int_\Omega \phi_i \mathcal{B} \phi_j \, dx, \quad i, j \in \{1, \dots, n\}.$$

2.3.4 Finite-dimensional approximation of the prior

Next, we derive the finite-dimensional representation of the prior. The matrix representation of the operator $\mathcal{A}$ defined in Section 2.2.3 is given by the stiffness matrix $\boldsymbol{K}$ with entries

$$K_{ij} = \alpha \int_\Omega (\boldsymbol{\Theta}(\boldsymbol{x}) \nabla \phi_i(\boldsymbol{x})) \cdot \nabla \phi_j(\boldsymbol{x}) + \phi_i(\boldsymbol{x}) \phi_j(\boldsymbol{x}) \, d\boldsymbol{x}, \quad i, j \in \{1, \dots, n\}.$$

It follows that both $\boldsymbol{A} = \boldsymbol{M}^{-1}\boldsymbol{K}$ and $\boldsymbol{A}^{-1} = \boldsymbol{K}^{-1}\boldsymbol{M}$ are self-adjoint operators in the sense of (2.13).

We are now in a position to define the finite-dimensional Gaussian prior measure μ_0^h specified by the following density (with respect to the Lebesgue measure):

$$\pi_{\text{prior}}(\boldsymbol{m}) \propto \exp\left[-\frac{1}{2} \|\boldsymbol{A}(\boldsymbol{m} - \boldsymbol{m}_0)\|_M^2\right]. \tag{2.17}$$

This definition implies that $\boldsymbol{\Gamma}_{\text{prior}} = \boldsymbol{A}^{-2}$.

2.3.5 Finite-dimensional approximation of the posterior

In infinite dimensions, the Bayes' formula (2.5) has to be expressed in terms of the Radon–Nikodym derivative since the prior and posterior distri-

butions do not have density functions with respect to the Lebesgue measure. Since we approximate the prior measure μ_0 by μ_0^h, it is natural to define a finite-dimensional approximation $\mu^{y,h}$ of the posterior measure μ^y such that

$$\frac{d\mu^{y,h}}{d\mu_0^h} = \frac{1}{Z^h}\pi_{\text{like}}(\boldsymbol{y}^{\text{obs}}|m_h),$$

where $Z^h = \int_X \pi_{\text{like}}(\boldsymbol{y}^{\text{obs}}|\boldsymbol{m})\, d\mu_0^h$, and π_{like} is the likelihood (2.4) evaluated at m_h. If we define $\pi_{\text{post}}(\boldsymbol{m}|\boldsymbol{y}^{\text{obs}})$ as the density of $\mu^{y,h}$, again with respect to the Lebesgue measure, we recover the familiar finite-dimensional Bayes' formula

$$\pi_{\text{post}}(\boldsymbol{m}|\boldsymbol{y}^{\text{obs}}) \;\propto\; \pi_{\text{prior}}(\boldsymbol{m})\pi_{\text{like}}(\boldsymbol{y}^{\text{obs}}|m_h), \tag{2.18}$$

where the normalization constant $1/Z^h$, which does not depend on $\boldsymbol{m}$, is omitted. Finally, we can express the posterior pdf explicitly as

$$\pi_{\text{post}}(\boldsymbol{m}) \;\propto\; \exp\left(-\frac{1}{2}\left\|\boldsymbol{f}(m_h) - \boldsymbol{y}^{\text{obs}}\right\|^2_{\boldsymbol{\Gamma}^{-1}_{\text{noise}}} - \frac{1}{2}\left\|\boldsymbol{A}(\boldsymbol{m}-\boldsymbol{m}_0)\right\|^2_{\boldsymbol{M}}\right), \tag{2.19}$$

where, to recall our notation, $m_h = \sum_{j=1}^n m_j\phi_j \in V_h$ and $\boldsymbol{m} = (m_1, \ldots, m_n)^T$. We observe that the negative log of the right side of (2.19) is the finite-dimensional approximation of the objective functional in (2.8).

As a finite-dimensional counterpart of Section 2.2.5, we linearize the parameter-to-observable map $\boldsymbol{f}$ at the MAP point, but now considering it as a function of the coefficient vector $\boldsymbol{m}$. Let $\boldsymbol{\Gamma}_{\text{post}}$ be the posterior covariance matrix in the $\boldsymbol{M}$-inner product. Using (2.10), we obtain

$$\boldsymbol{\Gamma}_{\text{post}} = \left(\boldsymbol{F}^\natural\boldsymbol{\Gamma}^{-1}_{\text{noise}}\boldsymbol{F} + \boldsymbol{\Gamma}^{-1}_{\text{prior}}\right)^{-1}, \tag{2.20}$$

with $\boldsymbol{F}^\natural = \boldsymbol{M}^{-1}\boldsymbol{F}^T$ as defined in (2.14). Note that $\boldsymbol{\Gamma}_{\text{post}}$ is self-adjoint, i.e., $\boldsymbol{\Gamma}_{\text{post}} = \boldsymbol{\Gamma}_{\text{post}}^*$ in the sense of (2.13).

Since the posterior covariance matrix $\boldsymbol{\Gamma}_{\text{post}}$ is typically dense, we wish to avoid explicitly storing it, especially when the parameter dimension n is large. Even if we are able to do so, it is prohibitively expensive to construct. The reason is that the Jacobian of the parameter-to-observable map, $\boldsymbol{F}$, is generally a dense matrix, and its construction typically requires n forward PDE solves. This is clearly intractable when n is large and solving the PDEs is expensive. However, one can exploit the structure of the inverse problem, to approximate the posterior covariance matrix with desired accuracy, as we shall show in Section 2.5.

2.3.6 Sample generation in a finite element discretization

We begin by developing expressions for a general Gaussian distribution with mean $\bar{\boldsymbol{m}}$ and covariance matrix $\boldsymbol{\Gamma}$. Then, they are specified for the Gaussian prior with $(\boldsymbol{m}_0, \boldsymbol{\Gamma}_{\text{prior}})$. Realizations of a finite-dimensional Gaussian random variable with mean $\bar{\boldsymbol{m}}$ and covariance matrix $\boldsymbol{\Gamma}$ can be found by choosing a vector $\boldsymbol{n}$ containing independent and identically distributed (*i.i.d.*) standard normal random values and computing

$$\boldsymbol{m} = \bar{\boldsymbol{m}} + \boldsymbol{L}\boldsymbol{n}, \tag{2.21}$$

where $\boldsymbol{L}$ is a linear map from $\mathbb{R}^n$ to $\mathbb{R}^n_{\boldsymbol{M}}$ such that $\boldsymbol{\Gamma} = \boldsymbol{L}\boldsymbol{L}^\diamond$, in which the adjoint $\boldsymbol{L}^\diamond = \boldsymbol{L}^T\boldsymbol{M}$ (see also (2.15)). Note that $\boldsymbol{M}^{-1/2}\boldsymbol{n}$ is a sample from $\mathcal{N}(\boldsymbol{0}, \boldsymbol{I})$ in the mass-weighted inner product.

In particular, for $\bar{\boldsymbol{m}} = \boldsymbol{m}_0$ and $\boldsymbol{\Gamma} = \boldsymbol{\Gamma}_{\text{prior}}$, we have $\boldsymbol{L}_{\text{prior}} = \boldsymbol{K}^{-1}\boldsymbol{M}\boldsymbol{M}^{-1/2} = \boldsymbol{K}^{-1}\boldsymbol{M}^{1/2}$ (see Appendix I) and samples from the prior are computed as $\boldsymbol{m} = \boldsymbol{m}_0 + \boldsymbol{K}^{-1}\boldsymbol{M}^{1/2}\boldsymbol{n}$. Samples from the Gaussian approximation to the posterior use the low-rank representation introduced in Section 2.5 and the corresponding expressions are given in (2.30) and (2.31).

2.3.7 The pointwise variance field in a finite element discretization

Let us approximate the covariance function in V_h for a generic Gaussian measure with covariance operator $\mathcal{C}$. Recall from Section 2.2.3 that the covariance function $c(\boldsymbol{x}, \boldsymbol{y})$ corresponding to the covariance operator $\mathcal{C}$ is the Green's function of $\mathcal{C}^{-1}$, i.e.,

$$\mathcal{C}^{-1} c(\boldsymbol{x}, \boldsymbol{y}) := \delta_{\boldsymbol{y}}(\boldsymbol{x}) \text{ for } \boldsymbol{x} \in \Omega,$$

where $\delta_{\boldsymbol{y}}$ denotes the Dirac delta function concentrated at $\boldsymbol{y} \in \Omega$. We approximate $c(\boldsymbol{x}, \boldsymbol{y})$ in the finite element space V_h by $c_h(\boldsymbol{x}, \boldsymbol{y}) = \sum_{i=1}^{n} c_i(\boldsymbol{y})\, \phi_i(\boldsymbol{x})$ with coefficient vector $\boldsymbol{c}(\boldsymbol{y}) = [c_1(\boldsymbol{y}), \ldots, c_n(\boldsymbol{y})]^T$. Using the Galerkin finite element method to obtain a finite element approximation of the preceding equation results in

$$\mathbf{C}^{-1}\boldsymbol{c}(\boldsymbol{y}) = \boldsymbol{\Phi}(\boldsymbol{y}) \quad \text{with} \quad \boldsymbol{\Phi}(\boldsymbol{y}) = [\phi_1(\boldsymbol{y}), \ldots, \phi_n(\boldsymbol{y})]^T$$

and the entries of the matrix $\mathbf{C}^{-1}$ are given by $C_{ij}^{-1} = (\phi_i, \mathcal{C}^{-1}\phi_j)_{L^2(\Omega)}$. It follows that $\boldsymbol{c}(\boldsymbol{y}) = \mathbf{C}\boldsymbol{\Phi}(\boldsymbol{y})$ and

$$c_h(\boldsymbol{x}, \boldsymbol{y}) = \sum_{i=1}^{n} c_i(\boldsymbol{y})\Phi_i(\boldsymbol{x}) = \boldsymbol{\Phi}(\boldsymbol{x})^T \mathbf{C}\boldsymbol{\Phi}(\boldsymbol{y}) \quad \text{for } \boldsymbol{x}, \boldsymbol{y} \in \Omega.$$

Let us denote by $\mathbf{\Gamma}^{-1}$ the representation of $P_h \mathcal{C}^{-1} P_h'$ in V_h; then, using (2.16) yields that $\mathbf{C} = \mathbf{\Gamma} \boldsymbol{M}^{-1}$. Consequently, the discretized covariance function for the covariance operator $\mathcal{C}$ now becomes

$$c_h(\boldsymbol{x}, \boldsymbol{y}) = \mathbf{\Phi}(\boldsymbol{x})^T \mathbf{\Gamma} \boldsymbol{M}^{-1} \mathbf{\Phi}(\boldsymbol{y}). \tag{2.22}$$

Let us now apply (2.22) to compute the prior variance field. As discussed in Section 2.3.4, $\mathbf{\Gamma}_{\text{prior}} = \boldsymbol{A}^{-2} = \boldsymbol{K}^{-1} \boldsymbol{M} \boldsymbol{K}^{-1} \boldsymbol{M}$. This results in the discretized prior covariance function

$$c_h^{\text{prior}}(\boldsymbol{x}, \boldsymbol{y}) = \mathbf{\Phi}(\boldsymbol{x})^T \boldsymbol{K}^{-1} \boldsymbol{M} \boldsymbol{K}^{-1} \mathbf{\Phi}(\boldsymbol{y}),$$

By taking $\boldsymbol{y} = \boldsymbol{x}$, the prior variance field at an arbitrary point $\boldsymbol{x} \in \Omega$ reads

$$c_h^{\text{prior}}(\boldsymbol{x}, \boldsymbol{x}) = \mathbf{\Phi}(\boldsymbol{x})^T \boldsymbol{K}^{-1} \boldsymbol{M} \boldsymbol{K}^{-1} \mathbf{\Phi}(\boldsymbol{x}).$$

The pointwise variance field of the posterior distribution builds on the low-rank representation introduced in Section 2.5. The resulting expression, which requires the prior variance field, is given in (2.29).

2.4 Finding the MAP point

Section 2.2.4 introduced the idea of the MAP point as a first step in exploring the solution of the statistical inverse problem. To find the MAP point, one needs to solve a discrete approximation (using the discretizations of Section 2.3) of the optimization problem (2.8), which amounts to a large-scale nonlinear least squares numerical optimization problem. In this section,

we provide just a brief summary of a scalable method we use for solving this problem, and refer the reader to our earlier work for details, in particular the work on inverse wave propagation [3, 4, 67]. We use an inexact matrix-free Newton-conjugate gradient (CG) method in which only Hessian-vector products are required. These Hessian-vector products are computed by solving a linearized forward-like and an adjoint-like PDE problems, and thus the Hessian matrix is never constructed explicitly. Inner CG iterations are terminated prematurely when sufficient reduction is made in the norm of the gradient, or when a direction of negative curvature is encountered. The prior operator is used to precondition the CG iterations. Globalization is through an Armijo backtracking line search.

Because the major components of the method can be expressed as solving PDE-like systems, the method inherits the scalability (with respect to problem dimension) of the forward PDE solve. The remaining ingredient for overall scalability is that the optimization algorithm itself be scalable with increasing problem size. This is indeed the case: for a wide class of nonlinear inverse problems, the outer Newton iterations and the inner CG iterations are independent of the mesh size, as is found to be the case for instance for inverse wave propagation [4, 67]. This is a consequence of the use of a Newton solver, of the compactness of the Hessian of the data misfit term (i.e., the first term) in (2.9), and of the use of preconditioning by $\mathbf{\Gamma}_{\text{prior}}$, so that the resulting preconditioned Hessian is a compact perturbation of the identity, for which CG exhibits mesh-independent iterations.

2.5 Low rank approximation of the Hessian matrix

2.5.1 Overview

As discussed in Section 2.2.5, linearizing the parameter-to-observable map results in the posterior covariance matrix being given by the inverse of the Hessian of the negative log posterior. Explicitly computing this Hessian matrix requires a (linearized) forward PDE problem for each of its columns, and thus as many (linearized) forward PDE solves are required as there are parameters. For inverse problems in which one seeks to infer an unknown parameter field, discretization results in a very large number of parameters; explicitly computing the Hessian—and hence the covariance matrix—is thus out of the question. As a remedy, we exploit the structure of the problem to find an approximation of the Hessian that can be constructed and dealt with efficiently.

When the linearized parameter-to-observable map is used in $\mathcal{J}(m)$ (as defined in (2.9)) and second derivatives of the resulting functional are computed, one obtains the Gauss-Newton portion of the Hessian of $\mathcal{J}(m)$. Both, the full Hessian matrix as well as its Gauss Newton portion are positive definite at the MAP point and they only differ in terms that involve the adjoint variable. Since the adjoint system is driven only by the data misfit (see, for instance, the adjoint wave equation 2.36), the adjoint variable is expected to be small when the data misfit is small, which occurs provided the model and observational errors are not too large. The Gauss-Newton portion of the Hessian is thus often a good approximation of the full Hessian of $\mathcal{J}(m)$.

For conciseness and convenience of the notation, we focus on computing a low rank approximation of the Gauss-Newton portion of the (misfit) Hessian in Section 2.5.2. The same approach also applies to the computation of a low rank approximation of the full Hessian, whose inverse might be a better approximation for the covariance matrix if the data is very noisy and the data misfit at the MAP point cannot be neglected. The low rank construction of the misfit Hessian is based on the Lanczos method and thus only requires Hessian-vector products. Using the Sherman-Morrison-Woodbury formula, this approximation translates into an approximation of the posterior covariance matrix.

In Section 2.5.3, we present low rank-exploiting methods for sample generation from the Gaussian approximation of the posterior, as well as methods for the efficient computation of the pointwise variance field. Finally, in Section 2.5.4, we discuss the overall scalability of our approach.

2.5.2 Low rank covariance approximation

For many ill-posed inverse problems, the Gauss-Newton portion of the Hessian matrix (called the Gauss-Newton Hessian for short) of the data misfit term in (2.9) evaluated at any $\mathbf{m}$,

$$\boldsymbol{H}_{\text{misfit}} := \boldsymbol{F}^{\natural}\boldsymbol{\Gamma}_{\text{noise}}^{-1}\boldsymbol{F}, \tag{2.23}$$

behaves like (the discretization of) a compact operator (see, e.g., [198, p.17]). The intuitive reason for this is that only parameter modes that strongly influence the observations through the linearized parameter-to-observable map $\boldsymbol{F}$

will be present in the dominant spectrum of the Hessian (2.23). For typical inverse problems, observations are sparse, and hence the dimension of the observable space is much smaller than that of the parameter space. Furthermore, highly oscillatory perturbations in the parameter field often have negligible effect on the output of the parameter-to-observable map. In [29, 30], we have shown that the Gauss-Newton Hessian of the data misfit is a compact operator, and that for smooth media its eigenvalues decay exponentially to zero. Thus, the range space of the Gauss-Newton Hessian is effectively finite-dimensional even before discretization, i.e., it is independent of the mesh. We can exploit the compact nature of the data misfit Hessian to construct scalable algorithms for approximating the inverse of the Hessian [73, 143].

A simple manipulation of (2.20) yields the following expression for the posterior covariance matrix, which will prove convenient for our purposes:

$$\boldsymbol{\Gamma}_{\text{post}} = \boldsymbol{\Gamma}_{\text{prior}}^{1/2} \left(\boldsymbol{\Gamma}_{\text{prior}}^{1/2} \boldsymbol{H}_{\text{misfit}} \boldsymbol{\Gamma}_{\text{prior}}^{1/2} + \boldsymbol{I} \right)^{-1} \boldsymbol{\Gamma}_{\text{prior}}^{1/2}. \tag{2.24}$$

We now present a fast method for approximating $\boldsymbol{\Gamma}_{\text{post}}$ with controllable accuracy by making a low rank approximation of the so-called *prior-preconditioned Hessian of the data misfit*, namely,

$$\tilde{\boldsymbol{H}}_{\text{misfit}} := \boldsymbol{\Gamma}_{\text{prior}}^{1/2} \boldsymbol{H}_{\text{misfit}} \boldsymbol{\Gamma}_{\text{prior}}^{1/2}. \tag{2.25}$$

Let $(\lambda_i, \boldsymbol{v}_i)\,, i = 1, \ldots, n$ be the eigenpairs of $\tilde{\boldsymbol{H}}_{\text{misfit}}$, and $\boldsymbol{\Lambda} = \text{diag}(\lambda_1, \ldots, \lambda_n) \in \mathbb{R}^{n \times n}$. Define $\boldsymbol{V} \in \mathbb{R}^{n \times n}$ such that its columns are the eigenvectors $\boldsymbol{v}_i$ of $\tilde{\boldsymbol{H}}_{\text{misfit}}$. Replacing $\tilde{\boldsymbol{H}}_{\text{misfit}}$ by its spectral decomposition (recall that $\boldsymbol{V}^\diamond$ is the adjoint

of $\boldsymbol{V}$ as defined in (2.15)),

$$\left(\boldsymbol{\Gamma}_{\text{prior}}^{1/2}\boldsymbol{H}_{\text{misfit}}\boldsymbol{\Gamma}_{\text{prior}}^{1/2}+\boldsymbol{I}\right)^{-1}=(\boldsymbol{V}\boldsymbol{\Lambda}\boldsymbol{V}^{\diamond}+\boldsymbol{I})^{-1}.$$

When the eigenvalues of $\tilde{\boldsymbol{H}}_{\text{misfit}}$ decay rapidly we can construct a low-rank approximation of $\tilde{\boldsymbol{H}}_{\text{misfit}}$ by computing only the r largest eigenvalues, i.e.,

$$\boldsymbol{\Gamma}_{\text{prior}}^{1/2}\boldsymbol{H}_{\text{misfit}}\boldsymbol{\Gamma}_{\text{prior}}^{1/2}=\boldsymbol{V}_r\boldsymbol{\Lambda}_r\boldsymbol{V}_r^{\diamond}+\mathcal{O}\left(\sum_{i=r+1}^{n}\lambda_i\right),$$

where $\boldsymbol{V}_r\in\mathbb{R}^{n\times r}$ contains r eigenvectors of $\tilde{\boldsymbol{H}}_{\text{misfit}}$ corresponding to the r largest eigenvalues $\lambda_i, i=1,\ldots,r$, and $\boldsymbol{\Lambda}_r=\text{diag}(\lambda_1,\ldots,\lambda_r)\in\mathbb{R}^{r\times r}$. We can then invert the approximate Hessian using the Sherman-Morrison-Woodbury formula to obtain

$$\left(\boldsymbol{\Gamma}_{\text{prior}}^{1/2}\boldsymbol{H}_{\text{misfit}}\boldsymbol{\Gamma}_{\text{prior}}^{1/2}+\boldsymbol{I}\right)^{-1}=\boldsymbol{I}-\boldsymbol{V}_r\boldsymbol{D}_r\boldsymbol{V}_r^{\diamond}+\mathcal{O}\left(\sum_{i=r+1}^{n}\frac{\lambda_i}{\lambda_i+1}\right),\qquad(2.26)$$

where $\boldsymbol{D}_r:=\text{diag}(\lambda_1/(\lambda_1+1),\ldots,\lambda_r/(\lambda_r+1))\in\mathbb{R}^{r\times r}$. Equation (2.26) shows the truncation error due to the low-rank approximation based on the first r eigenvalues. To obtain an accurate approximation of $\boldsymbol{\Gamma}_{\text{post}}$, only eigenvectors corresponding to eigenvalues that are small compared to 1 can be neglected. With such a low-rank approximation, the final expression for the approximate posterior covariance is given by

$$\boldsymbol{\Gamma}_{\text{post}}\approx\boldsymbol{\Gamma}_{\text{prior}}-\boldsymbol{\Gamma}_{\text{prior}}^{1/2}\boldsymbol{V}_r\boldsymbol{D}_r\boldsymbol{V}_r^{\diamond}\boldsymbol{\Gamma}_{\text{prior}}^{1/2}.\qquad(2.27)$$

Note that (2.27) expresses the posterior uncertainty (in terms of the covariance matrix) as the prior uncertainty less any information gained from the

data. Due to the square root of the prior in the rightmost term in (2.27), the information gained from the data is filtered through the prior, i.e., only information consistent with the prior can reduce the posterior uncertainty.

2.5.3 Fast generation of samples and the pointwise variance field

Properties of the last term in (2.27), such as its diagonal (which provides the reduction in variance due to the knowledge acquired from the data) can be obtained numerically through just r applications of the square root of the prior covariance matrix to r columns of $\boldsymbol{V}_r$. Let us define

$$\tilde{\boldsymbol{V}}_r = \boldsymbol{\Gamma}_{\text{prior}}^{1/2} \boldsymbol{V}_r,$$

then (2.27) becomes

$$\boldsymbol{\Gamma}_{\text{post}} \approx \boldsymbol{\Gamma}_{\text{prior}} - \tilde{\boldsymbol{V}}_r \boldsymbol{D}_r \tilde{\boldsymbol{V}}_r^{\diamond}, \tag{2.28}$$

with $\tilde{\boldsymbol{V}}_r^{\diamond} = \boldsymbol{V}_r^{\diamond} \boldsymbol{\Gamma}_{\text{prior}}^{1/2}$.

The linearized posterior is a Gaussian distribution with known mean, namely the MAP point, and low rank-based covariance (2.28). Thus, the pointwise variance field and samples can be generated as in Section 2.3.7 and 2.3.6, respectively. The variance field can be computed as

$$c_h^{\text{post}}(\boldsymbol{x}, \boldsymbol{x}) = c_h^{\text{prior}}(\boldsymbol{x}, \boldsymbol{x}) - \sum_{k=1}^{r} d_k \left(\tilde{\boldsymbol{v}}_{kh}(\boldsymbol{x}) \right)^2, \tag{2.29}$$

where $\tilde{\boldsymbol{v}}_{kh}(\boldsymbol{x}) = \boldsymbol{\Phi}(\boldsymbol{x})^T \tilde{\boldsymbol{v}}_k$, with $\tilde{\boldsymbol{v}}_k$ denoting the kth column of $\tilde{\boldsymbol{V}}_r$, is the function in V_h corresponding to the nodal vector $\tilde{\boldsymbol{v}}_k$.

Now, we can compute samples from the posterior provided that we have a factorization $\boldsymbol{\Gamma}_{\text{post}} = \boldsymbol{L}\boldsymbol{L}^{\diamond}$. One possibility for $\boldsymbol{L}$ (see Appendix I for the detailed derivation) reads

$$\boldsymbol{L} := \boldsymbol{\Gamma}_{\text{prior}}^{1/2} \left(\boldsymbol{V}_r \boldsymbol{P}_r \boldsymbol{V}_r^{\diamond} + \boldsymbol{I} \right) \boldsymbol{M}^{-1/2} \tag{2.30}$$

with $\boldsymbol{P}_r = \text{diag}\left(1/\sqrt{\lambda_1 + 1} - 1, \ldots, 1/\sqrt{\lambda_r + 1} - 1\right) \in \mathbb{R}^{r \times r}$, $\boldsymbol{L}$ as a linear map from $\mathbb{R}^n$ to $\mathbb{R}^n_M$, and $\boldsymbol{I}$ as the identity map in both $\mathbb{R}^n$ and $\mathbb{R}^n_{\boldsymbol{M}}$. As discussed in Section 2.3.6, samples can be then computed as

$$\boldsymbol{\nu}^{\text{post}} = \boldsymbol{m}_{\text{MAP}} + \boldsymbol{L}\boldsymbol{n}, \tag{2.31}$$

where $\boldsymbol{n}$ is an *i.i.d.* standard normal random vector.

2.5.4 Scalability

We now discuss the scalability of the above low rank construction of the posterior covariance matrix in (2.27). The dominant task is the computation of the dominant spectrum of the prior preconditioned Hessian of the data misfit, $\tilde{\boldsymbol{H}}_{\text{misfit}}$, given by (2.25). Computing the spectrum by a matrix-free eigensolver such as Lanczos means that we need only form actions of $\tilde{\boldsymbol{H}}_{\text{misfit}}$ with a vector. As argued at the end of Section 2.3.5, the linearized parameter-to-observable map $\boldsymbol{F}$ is too costly to be constructed explicitly since it requires n linearized forward PDE solves. However, its action on a vector can be computed by solving a single linearized forward PDE (which we term the *incremental forward problem*), regardless of the number of parameters n and observations q. Similarly, the action of $\boldsymbol{F}^{\natural}$ on a vector can be found by solving a single

linearized adjoint PDE (which we term the *incremental adjoint problem*). Explicit expressions for the incremental forward and incremental adjoint PDEs in the context of inverse acoustic wave propagation will be given in Section 2.6. Solvers for the incremental forward and adjoint problems of course inherit the scalability of the forward PDE solver. The other major cost in computing the action of $\tilde{\boldsymbol{H}}_{\text{misfit}}$ on a vector is the application of the square root of the prior, $\boldsymbol{\Gamma}_{\text{prior}}^{1/2}$, to a vector. As discussed in Section 2.2.3, this amounts to solving a Laplacian-like problem. Using a scalable elliptic solver such as multigrid renders this component scalable as well. Therefore, the scalability of the application of $\tilde{\boldsymbol{H}}_{\text{misfit}}$ to a vector—which is the basic operation of a matrix-free eigenvalue solver such as Lanczos—is assured, and the cost is independent of the parameter dimension.

The remaining requirement for independence of parameter dimension in the construction of the low rank-based representation of the posterior covariance in (2.27) is that the number of dominant eigenvalues of $\boldsymbol{H}_{\text{misfit}}$ be independent of the dimension of the discretized parameter. This is the case when $\boldsymbol{H}_{\text{misfit}}$ and $\boldsymbol{\Gamma}_{\text{prior}}$ in (2.23) are discretizations of a compact and a continuous operator, respectively. The continuity of $\mathcal{C}_0$ is a direct consequence of the prior Gaussian measure μ_0; in fact, $\mathcal{C}_0$, the infinite-dimensional counterpart of $\boldsymbol{\Gamma}_{\text{prior}}$, is also a compact operator. Compactness of the data misfit Hessian $\boldsymbol{H}_{\text{misfit}}$ for inverse wave propagation problems has long been observed (e.g., [51]) and, as mentioned above, has been proved for frequency-domain acoustic inverse scattering for both continuous and pointwise observation op-

erators [29, 30]. Specifically, we have shown that the data misfit Hessian is a compact operator at any point in the parameter domain. We also quantify the decay of the data misfit Hessian eigenvalues in terms of the smoothness of the medium, i.e., the smoother it is the faster the decay rate. For an analytic target medium, the rate can be shown to be exponential. That is, the data misfit Hessian can be approximated well with a small number of its dominant eigenvectors and eigenvalues.

As a result, the number of Lanczos iterations required to obtain a low rank approximation of $\tilde{\boldsymbol{H}}_{\text{misfit}}$ is independent of the dimension of the discretized parameter field. Once the low-rank approximation of $\tilde{\boldsymbol{H}}_{\text{misfit}}$ is constructed, no additional forward or adjoint PDE solves are required. Any action of $\boldsymbol{\Gamma}_{\text{post}}$ in (2.27) on a vector (which is required to generate samples from the posterior distribution and to compute the diagonal of the covariance) is now dominated by the action of $\boldsymbol{\Gamma}_{\text{prior}}$ on a vector. But as discussed above, this amounts to an elliptic solve and can be readily carried out in a scalable manner. Since r is independent of the dimension of the discretized parameter field, estimating the posterior covariance matrix requires a constant number of forward/adjoint PDE solves, independent of the number of parameters, observations, and state variables. Moreover, since the dominant cost is that of solving forward and adjoint PDEs as well as elliptic problems representing the prior, scalability of the overall uncertainty quantification method follows when the forward and adjoint PDE solvers are scalable.

2.6 Application to global seismic statistical inversion

In this section, we apply the computational framework developed in the previous sections to the statistical inverse problem of global seismic inversion, in which we seek to reconstruct the heterogeneous compressional (acoustic) wave speed from observed seismograms, i.e., seismic waveforms recorded at points on earth's surface. With the rapid advances in observational capabilities, exponential growth in supercomputing, and maturation of forward seismic wave propagation solvers, there is great interest in solving the global seismic inverse problem governed by the full acoustic or elastic wave equations [70,165]. Already, successful deterministic inversions have been carried out at regional scales; for example, see [71, 72, 130, 191, 206].

We consider global seismic model problems in which the seismic source is taken as a simple point source. Sections 2.6.1 and 2.6.2 define the prior mean and covariance operator for the wave speed and its discretization. Section 2.6.3 presents the parameter-to-observable map $\boldsymbol{f}(m)$ (which involves solution of the acoustic wave equation) and the likelihood model. We next provide the expressions for the gradient and application of the Hessian of the negative log-likelihood in Section 2.6.4. Then, we discuss the discretization of the forward and adjoint wave equations and implementation details in Section 2.6.5. Section 2.6.6 provides the inverse problem setup, while numerical results and discussion are provided in Sections 2.6.7 and 2.6.8.

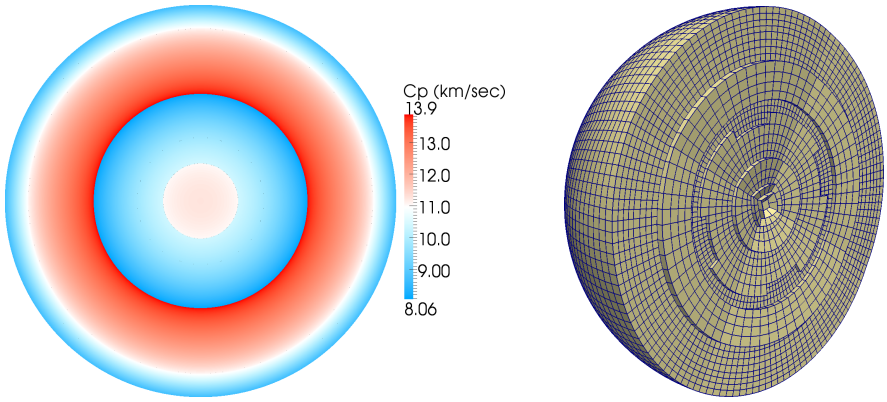

Figure 2.1: Left image: Cross section through the spherically symmetric PREM earth model, which is the prior mean in the inversion. Right image: Mesh used for both wave speed parameters (discretized with $N = 1$) and wave propagation unknowns ($N = 3$). The mesh is tailored to the local wave lengths.

2.6.1 Parameter space for seismic inversion

We are interested in inferring the heterogenous compressional acoustic wavespeed in the earth. In order to do this, we represent the earth as a sphere of radius 6,371km. We employ two earth models, i.e.; two representations of the compressional wave speed and density in the earth. We suppose that our current knowledge of the earth is given by the spherically symmetric Preliminary Reference Earth Model (PREM) [63], which is depicted in Figure 2.1.

The PREM is used as the mean of the prior distribution, and as the starting point for the determination of the MAP point by the optimization solver. Then, we presume that the real earth behaves according to the S20RTS velocity model [173], which superposes lateral wave speed variations on the

laterally-homogeneous PREM. S20RTS is used to generate waveforms used as synthetic observational data for inversion; we refer to it as the "ground truth" earth model. The inverse problem then aims to reconstruct the S20RTS ground truth model from the (noisy) synthetic data and from prior knowledge of the PREM, and quantify the uncertainty in doing so.

The parameter field m of interest for the inverse problem is the deviation or anomaly from the PREM, and hence it is sensible to choose the zero funtion as the prior mean. Owing to the prior covariance operator specified in Section 2.2.3, the deviation is smooth; in fact it is continuous almost surely. The wave speed parameter space is discretized using continuous isoparametric trilinear finite elements on a hexahedral octree-based mesh. To generate the mesh, we partition the earth into 3 layers described by 13 mapped cubes. The first layer consists of a single cube surrounded by two layers of six mapped cubes. The resulting mesh is aligned with the interface between the outer core and the mantle, where the wave speed has a significant discontinuity (see Figure 2.1). It is also aligned with several known weaker discontinuities between layers.

The parameter mesh coincides with the mesh used to solve the wave equation described in Section 2.6.3. The mesh is locally refined to resolve the local seismic wavelength resulting from a given frequency of interest for the PREM. We choose a conservative number of grid points per wavelength to permit the same mesh to be used for anticipated variations in the earth model across the iterations needed to determine the MAP point. For the parallel mesh

generation and its distributed storage, we use fast forest-of-octree algorithms for scalable adaptive mesh refinement from the `p4est` library [36, 37].

2.6.2 The choice of prior

Since the prior is a Gaussian measure, it is completely specified by a mean function and a covariance operator. As discussed in Section 2.6.1, the prior distribution for the anomaly (the deviation of the acoustic wavespeed from that described by the PREM model) is naturally chosen to have zero mean. The choice of covariance operator for the prior distribution has to encode several important features. Recall that we specify the covariance operator via the precision operator $\mathcal{A}$ in Section 2.2.3. Therefore, the size of the variance about the zero mean is set by α, while the product $\alpha\Theta$ determines the correlation length of the prior Gaussian random field. We next specify the scalar α and the tensor Θ based on the following observations of models for the local wave speeds in the earth.

- Smoothness. The parameter field describes the *effective* local wavespeed, which, for a finite source frequency, depends on the local average of material parameters within a neighborhood of each point in space. This makes the effective wave speed mostly a smooth field. Note that the S20RTS-based target wave speed model (see [173]) is smooth.

- Prior variance. The deviation in this effective wave speed from the PREM model is believed to be within a few percent. Thus, we select

α such that the prior standard deviation is about 3.5%. The S20RTS target model has a maximal deviation from PREM of 7%.

- Anisotropy in the mantle. We further incorporate the prior belief that the compressional wave speed has a stronger variation in depth than in the lateral directions. We encode this anisotropic variation through Θ. In particular, we select Θ such that the anisotropy is strongest near the surface, and gradually becomes weaker with higher overall correlation length at larger depths. We observe that the S20RTS target model also obeys a similar anisotropy.

From the preceding observations and discussion, we choose $\alpha = 1.5 \cdot 10^{-2}$, while Θ is chosen to have the following form

$$\Theta = \beta \left(\boldsymbol{I}_3 - \theta(\boldsymbol{x}) \boldsymbol{x}\boldsymbol{x}^T \right) \quad \text{with } \theta(\boldsymbol{x}) := \begin{cases} \frac{1-\theta}{r\|\boldsymbol{x}\|^2} \left(2\|\boldsymbol{x}\| - \frac{1}{r}\|\boldsymbol{x}\|^2 \right) & \text{if } \|\boldsymbol{x}\| \neq 0 \\ 0 & \text{if } \|\boldsymbol{x}\| = 0, \end{cases} \tag{2.32}$$

where $\boldsymbol{I}_3$ is the 3×3 identity matrix, $r = 6,371$km is the earth radius, $\beta = 7.5 \cdot 10^{-3} r^2$, and $\theta = 4 \cdot 10^{-2}$. The choice $0 < \theta < 1$ introduces anisotropy in Θ such that the prior assumes longer correlation lengths in tangential than in radial directions, and the anisotropy decreases smoothly towards the center of the sphere. In Figure 2.2 we show several Green's functions for the precision operator $\mathcal{A}^2$, which illustrate this anisotropy. Figure 2.3 shows a slice through the $\pm 2\sigma$ fields, through samples from the prior and through the ground truth model, which is used to generate the synthetic seismograms. Note that close to the boundary, the standard deviation of the prior becomes larger. This

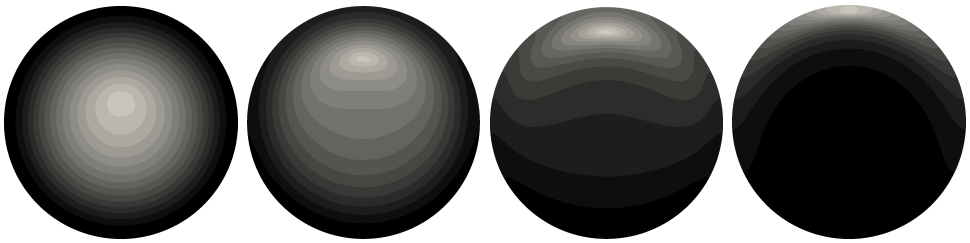

Figure 2.2: Contours of Green's functions at points in different depths for the precision operator of our prior $\mathcal{A}^2$. The contours are shown in slices through the Earth that contain the points, and larger values of the Green's function correspond to brighter shades of gray. These Green's functions correspond directly to the covariance function $c(\boldsymbol{x}, \boldsymbol{y})$ as discussed in Section 2.2. Note the anisotropy for points closer to the surface.

is partly a results of the anisotropy in the differential operator used in the construction of the prior, but mainly an effect of the homogeneous Neumann boundary condition used in the construction of the square root of the prior. This larger variance close to the boundary is also reflected in the prior samples, which have their largest values close to the boundary. Note that these samples have a larger correlation length in tangential than in normal directions, as intended by the choice of the anisotropy in (2.32). The ground truth model, which is also shown in Figure 2.3, is comparable to realizations of the prior in terms of magnitude as well as correlation.

2.6.3 The likelihood

In this section, we construct the likelihood (2.4) for the inverse acoustic wave problem. In order to do this, we need to construct the parameter-to-observable map $\boldsymbol{f}(m)$ and the observations $\boldsymbol{y}^{\text{obs}}$. Let us start by considering

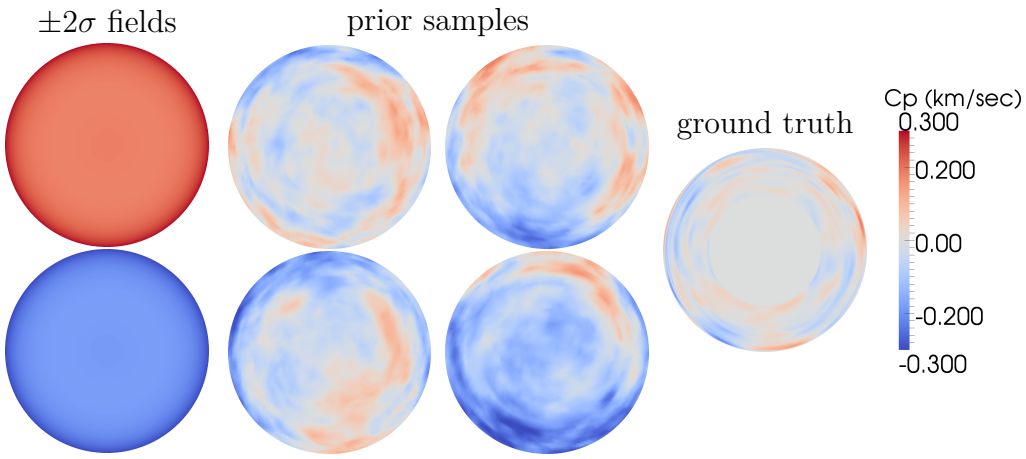

Figure 2.3: For illustration, we visualize several depictions of the prior using a common color scale. The images on the far left show slices through the pointwise positive and negative 2σ-deviation fields, which bound the pointwise 95% credible interval. The second and third columns show samples drawn from the prior distribution, while the fourth column depicts the "ground truth" parameter field. The prior has been chosen so that samples display similar qualitative features to the "ground truth" medium; they exhibit anisotropy in the outer layers of the mantle with larger correlation lengths in the lateral directions, and become more isotropic with higher overall correlation at depth.

the acoustic wave equation written in the first-order form,

$$\rho \frac{\partial \boldsymbol{v}}{\partial t} - \nabla(\rho c^2 e) = \boldsymbol{g}, \tag{2.33a}$$

$$\frac{\partial e}{\partial t} - \nabla \cdot \boldsymbol{v} = 0, \tag{2.33b}$$

where $\rho = \rho(\boldsymbol{x})$ denotes the mass density, $c = c(\boldsymbol{x})$ the local acoustic wave speed, $\boldsymbol{g}(\boldsymbol{x}, t)$ a (smoothed) point source $\boldsymbol{x} \in \Omega$, $\boldsymbol{v}(\boldsymbol{x}, t)$ the velocity, and $e(\boldsymbol{x}, t)$ the trace of the strain tensor, i.e., the dilatation. We equip (2.33) with the initial conditions

$$e(\boldsymbol{x}, 0) = e_0(\boldsymbol{x}), \text{ and } \boldsymbol{v}(\boldsymbol{x}, 0) = \boldsymbol{v}_0(\boldsymbol{x}), \quad \boldsymbol{x} \in \Omega, \tag{2.33c}$$

together with the boundary condition

$$e(\boldsymbol{x}, t) = 0, \quad \boldsymbol{x} \in \Gamma = \partial\Omega, t \in (0, T). \tag{2.33d}$$

Here, the acoustic wave initial-boundary value problem (2.33) is a simplified mathematical model for seismic waves propagation in the earth [6]. The choice of strain dilatation e together with velocity $\boldsymbol{v}$ in the first order system formulation is motivated from the strain-velocity formulation for the elastic wave equation used in [201].

Our goal is to quantify the uncertainty in inferring the spatially varying wave speed $m = c(\boldsymbol{x})$ from waveforms observed at receiver locations. To define the parameter-to-observable map for a given wave speed $c(\boldsymbol{x})$, we first solve the acoustic wave equation (2.33) given c, and record the velocity $\boldsymbol{v}$ at a finite number of receivers in the time interval $(0, T)$. Finally, we compute the Fourier

coefficients of the seismograms and truncate them; the truncated coefficients are the observables in the map $\boldsymbol{f}(m)$. A similar procedure is used to generate synthetic seismograms to define $\boldsymbol{y}^{\text{obs}}$. The noise covariance matrix $\boldsymbol{\Gamma}_{\text{noise}}^{-1}$ is prescribed as a diagonal matrix with constant variance.

2.6.4 Gradient and Hessian of the negative log posterior

Our proposed method for uncertainty quantification in Section 2.3 requires the computation of derivatives of the negative log posterior, which in turn requires the gradient and Hessian of the likelihood and the prior. These derivatives can be computed efficiently using an adjoint method, as we now show. For clarity, we derive the gradient and action of the Hessian in an infinite-dimensional setting. Let us begin by denoting $\boldsymbol{v}(c)$ as the space-time solution of the wave equation given the wave speed $c = c(\boldsymbol{x})$, and $\mathcal{B}$ as the observation operator. The parameter-to-observable map $\boldsymbol{f}(c)$ can be written as $\mathcal{B}\boldsymbol{v}(c)$. Thus, the negative log posterior is (compare with (2.9))

$$\mathcal{J}(c) := \frac{1}{2}\left\|\mathcal{B}\boldsymbol{v}(c) - \boldsymbol{y}^{\text{obs}}\right\|^2_{\boldsymbol{\Gamma}_{\text{noise}}^{-1}} + \frac{1}{2}\left\|\mathcal{A}(c - c_0)\right\|^2_{L^2(\Omega)}, \tag{2.34}$$

where $\boldsymbol{\Gamma}_{\text{noise}}$ is specified in Section 2.6.6. The dependence on the wave speed c of the velocity $\boldsymbol{v}$ and dilatation e is given by the solution of the forward wave propagation equation (2.33). The adjoint approach [67] allows us to write $\mathcal{G}(c)$, the gradient of $\mathcal{J}$ at a point c in parameter space, as

$$\mathcal{G}(c) := 2\rho c \int_0^T e(\nabla \cdot \boldsymbol{w})\, dt + \mathcal{A}^2(c - c_0), \tag{2.35}$$

where the adjoint velocity $\boldsymbol{w}$ and adjoint strain dilatation d satisfy the *adjoint wave propagation terminal-boundary value problem*

$$
\begin{aligned}
-\rho\frac{\partial \boldsymbol{w}}{\partial t} + \boldsymbol{\nabla}(c^2\rho d) &= -\mathcal{B}^*\boldsymbol{\Gamma}_{\text{noise}}^{-1}(\mathcal{B}\boldsymbol{v} - \boldsymbol{y}^{\text{obs}}) && \text{in } \Omega\times(0,T), && \text{(2.36a)}\\
-\frac{\partial d}{\partial t} + \nabla\cdot\boldsymbol{w} &= 0 && \text{in } \Omega\times(0,T), && \text{(2.36b)}\\
\rho\boldsymbol{w} = \boldsymbol{0}, d &= 0 && \text{in } \Omega\times\{t=T\}, && \text{(2.36c)}\\
d &= 0 && \text{on } \Gamma\times(0,T). && \text{(2.36d)}
\end{aligned}
$$

Here, $\mathcal{B}^*$, an operator from $\mathbb{R}^q$ to the space-time cylinder $\Omega\times(0,T)$, is the adjoint of $\mathcal{B}$. Note that the adjoint wave equations must be solved backward in time (due to final time data) and have the data misfit as a source term, but otherwise resemble the forward wave equations.

Similar to the computation of the gradient, the Hessian operator of $\mathcal{J}$ at c acting on an arbitrary variation $\tilde{c}$ is given by

$$
\mathcal{H}(c)\tilde{c} := 2\rho\int_0^T ce(\nabla\cdot\tilde{\boldsymbol{w}}) + c\tilde{e}(\nabla\cdot\boldsymbol{w}) + \tilde{c}e(\nabla\cdot\boldsymbol{w})\,dt + \mathcal{A}^2\tilde{c}, \qquad \text{(2.37)}
$$

where $\tilde{\boldsymbol{v}}$ and $\tilde{e}$ satisfy the *incremental forward wave propagation initial-boundary value problem*

$$
\begin{aligned}
\rho\frac{\partial \tilde{\boldsymbol{v}}}{\partial t} - \nabla(\rho c^2\tilde{e}) &= \nabla(2\rho c\tilde{c}e) && \text{in } \Omega\times(0,T),\\
\frac{\partial \tilde{e}}{\partial t} - \nabla\cdot\tilde{\boldsymbol{v}} &= 0 && \text{in } \Omega\times(0,T),\\
\rho\tilde{\boldsymbol{v}} = \boldsymbol{0}, \tilde{e} &= 0 && \text{in } \Omega\times\{t=0\},\\
\tilde{e} &= 0 && \text{on } \Gamma\times(0,T).
\end{aligned}
$$

On the other hand, $\tilde{\boldsymbol{w}}$ and $\tilde{d}$ satisfy the *incremental adjoint wave propagation terminal-boundary value problem*

$$\begin{aligned}
-\rho\frac{\partial\tilde{\boldsymbol{w}}}{\partial t}+\nabla(c^2\rho\tilde{d}) &= -\nabla(2\tilde{c}c\rho d)-\boldsymbol{\mathcal{B}}^*\boldsymbol{\Gamma}_{\text{noise}}^{-1}\boldsymbol{\mathcal{B}}\tilde{\boldsymbol{v}} && \text{in } \Omega\times(0,T),\\
-\frac{\partial\tilde{d}}{\partial t}+\nabla\cdot\tilde{\boldsymbol{w}} &= 0 && \text{in } \Omega\times(0,T),\\
\rho\tilde{\boldsymbol{w}}=\boldsymbol{0},\tilde{d} &= 0 && \text{in } \Omega\times\{t=T\},\\
\tilde{d} &= 0 && \text{on } \Gamma\times(0,T).
\end{aligned}$$

As can be seen, the incremental forward and incremental adjoint wave equations are linearizations of their forward and adjoint counterparts, and thus differ only in the source terms. Moreover, we observe that the computation of the gradient and the Hessian action amounts to solving a pair of forward/adjoint and a pair of incremental-forward/incremental-adjoint wave equations, respectively. For our computations, we use the Gauss-Newton approximation of the Hessian, which is guaranteed to be positive. This amounts to neglecting the terms that contain $\nabla\cdot\boldsymbol{w}$ in (2.37), and neglecting the term that includes d in the incremental adjoint wave equations.

2.6.5 Discretization of the wave equation and implementation details

We use the same hexahedral mesh to compute the wave solution $(\boldsymbol{v},e)$ as is used for the parameter c. While the parameter is discretized using trilinear finite elements, the wave equation, and its three variants (the adjoint, the incremental forward, and the incremental adjoint), are solved using a high-order discontinuous Galerkin (dG) method. The method, for which details

are provided in [28, 201], supports *hp*-non-conforming discretization, but only *h*-non-conformity is used in our implementation. For efficiency and scalability, a tensor product of Lagrange polynomials of degree N (we use $N \in \{2, 3, 4\}$ for the examples in the next section) is employed together with a collocation method based on Legendre-Gauss-Lobatto (LGL) nodes. As a result, the mass matrix is diagonal, which facilitates time integration using the classical four-stage fourth-order Runge Kutta method. We equip our dG method with exact Riemann numerical fluxes at element faces. To treat the non-conformity, we use the mortar approach of Kopriva [122, 123] to replace non-conforming faces by mortars that connect pairs of contributing elements. The actual computations are performed on the mortars instead of the non-conforming faces, and the results are then projected onto the contributing element faces. The method has been shown to be consistent, stable, convergent with optimal order, and highly scalable [28, 201].

It should be pointed out that the discretizations of the gradient and Hessian action given in Section 2.6.4 are not consistent with the discrete gradient and Hessian-vector product obtained by first discretizing the negative log posterior and then differentiating it. Here, inconsistency means that the former are equivalent to the latter only in the limit as the mesh size approaches zero (see also [93, 112]). The reason is that additional jump terms due to numerical fluxes at element interfaces are introduced in the discontinuous Galerkin discretization of the wave equation. In our implementation, we include these terms to ensure consistency, and this is verified by comparing the

discretized gradient and Hessian action expressions with their finite difference approximations.

Moreover, since we use a continuous Galerkin finite element method for the parameter, but a discontinuous Galerkin method for the wave solution, it is necessary to prolongate the parameter to the solution space before solving the forward wave equation, and its variants (adjoint, incremental state, incremental adjoint). Conversely, the gradient and the Hessian-vector application are computed in the wave solution space, and then restricted to the parameter space to provide the correct derivatives for the optimization solver. To ensure the symmetry of the Hessian, we construct these restriction and prolongation operations such that they are adjoint of each other.

Our discretization approach for the Bayesian inverse problem in Section 2.3 requires the repeated application of $\boldsymbol{A}^{-1}$, each amounting to an elliptic PDE solve on the finite-dimensional parameter space. To accomplish this task efficiently, we use the parallel algebraic multigrid (AMG) solver *ML* from the Trilinos project [79]. The cost of this elliptic solve is negligible compared to that of solving the time-dependent seismic wave equations, which employ high order discretization in contrast to the trilinear discretization of the anisotropic Poisson operator, $\boldsymbol{A}$.

The adjoint equation has to be solved backwards-in-time (as shown in Section 2.6.4); computation of the gradient (2.35) requires combinations of the state and adjoint solutions corresponding to same time. Thus the gradient computation requires the complete time history of the forward solve, which

cannot be stored due to the large-scale nature of our problem. A similar, but slightly more challenging storage problem occurs in the Hessian-vector application. Here, solving the incremental state equation requires the solution of the state equation, and the incremental adjoint solution requires the solution of the incremental state equation. We avoid storage of the time history of these wavefields by using a checkpointing method as employed in [67]. This scheme reduces the necessary storage at the expense of increasing the number of wave propagation solves.

Between 1200 and 4096 processor cores[4] for 10-20 hours are needed to solve the seismic inverse problems presented in the next section. The vast majority of the runtime is spent on computing solutions of the forward, adjoint and incremental wave equations either for the computation of the MAP point (see Section 2.4) or the Lanczos iterations for computing the low rank approximation of the misfit Hessian (see Section 2.6.7). Due to the large number of required wave propagation solves, good strong scalability of the wave propagation solver is important for rapid turnaround. We refer to the discussion in [26] on the scalability of the wave propagation solver, as well to the overall scalability of our Bayesian inversion approach applied to seismic inverse problems of up to one million parameters.

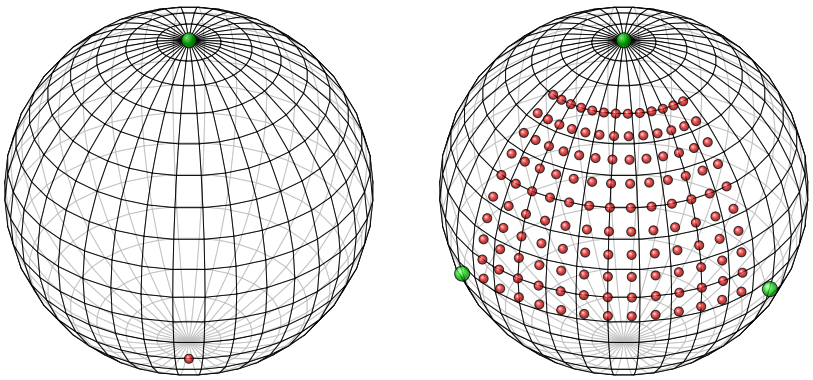

Figure 2.4: Location of sources (green) and receivers (red) for Problem I (left) and Problem II (right).

2.6.6 Setup of model problems

Synthetic observations $\boldsymbol{y}^{\text{obs}}$ are generated from solution of the wave equation using the S20RTS earth model. To mitigate the inverse crime [117], the local wave speed on LGL nodes of the wave propagation mesh is used to generate the observations, which implies that a higher order approximation of the earth model is used to generate the synthetic data, but the inversion is carried out on a lower order mesh. Both sources and receivers are located at 10km depth from the earth surface. For the source term $\boldsymbol{g}$ in (2.33), we use a delta function point source in the z-direction convolved with a narrow Gaussian in space. In time, we employ a Gaussian with standard deviation of $\sigma = 20$s centered at 60s. The wave propagation mesh (i.e., the discretization of velocity and dilatation) is chosen fine enough to accurately resolve frequencies below

[4]These computations were performed on the Texas Advanced Computing Center's Lonestar 4 system, which has 22,656 Westmere processor cores with 2GB memory per core.

0.05Hz. We Fourier transform the (synthetic) observed velocity waveforms at each receiver location and retain only the first 101 Fourier modes to define the observations $\boldsymbol{y}^{\text{obs}}$. In our problems, the Fourier coefficients $\boldsymbol{y}^{\text{obs}}$ vary between 10^{-5} and 10^{-1}, and we choose for the noise covariance a diagonal matrix with a standard deviation of 0.002.

We consider the following two model problems:

- **Problem I:** The first problem has a single source at the North pole and a single receiver at 45° south of the equator, as illustrated in the left image of Figure 2.4. The wave propagation time is 1800s. The wave speed (i.e., unknown material parameter) field is discretized on a mesh of trilinear hexahedra with 78,558 nodes, representing the unknowns in the inverse problem. The forward problem is discretized on the same mesh with 3rd-order dG elements, resulting in about 21.4 million spatial wave propagation unknowns, and in 2100 four-stage, fourth-order Runge Kutta time steps.

- **Problem II:** The second problem uses 130 receivers distributed on a quarter of the Northern hemisphere along zonal lines with 7.5° spacing and 3 simultaneous sources as shown on the right of Figure 2.4. The wave propagation time is 1200s. The wave speed is discretized on three different trilinear hexahedral meshes with 40,842, 67,770 and 431,749 wave speed parameters, which represent the unknowns in the inverse problem. These meshes corresponding to discretizations with 4th, 3rd

and 2nd order discontinuous elements for the wave propagation variables (velocity and dilatation). The results in the next section were computed with 67,770 wave speed parameters and the 3rd order dG discretization for velocity and dilatation. This amounts to 18.7 million spatial wave propagation unknowns, and 1248 Runge Kutta time steps.

2.6.7 Low rank approximation of the prior-preconditioned misfit Hessian

Before discussing the results for the quantification of the uncertainty in the solution of our inverse problems, we numerically study the spectrum of the prior-preconditioned misfit Hessian. In Figure 2.5, we show the dominant spectrum of the prior-preconditioned Hessian evaluated at the MAP estimate for Problem I (left) and Problem II (right). As can be observed, the eigenvalues decay faster in the former than in the latter. That is, the former is more ill-posed than the latter. The reason is that the three simultaneous source and 130 receivers of Problem II provide more information on the earth model. This implies that retaining more eigenvalues is necessary to accurately approximate the prior-preconditioned Hessian of the data misfit for Problem II compared to Problem I. In particular, we need at least 700 eigenvalues for Problem II as compared to about 40 for Problem I to obtain a sensible low-rank approximation of the Hessian, and this constitutes the bulk of computation time (since each Hessian-vector product in the Lanczos solver requires incremental forward and adjoint wave propagation solutions). These numbers compare with a total number of parameters of 78,558 (Problem I) and 67,770 (Problem II),

which amounts to a reduction of between two and three orders of magnitude. This directly translates into two to three orders of magnitude reduction in cost of solving the statistical inverse problem.

Figure 2.5 presents the spectra for Problem II for three different discretization of the wave speed parameter field. The figure suggests that the dominant spectrum is essentially mesh-independent and that all three parameter meshes are sufficiently fine to resolve the dominant eigenvectors of the prior-preconditioned Hessian. Consequently, the Hessian low-rank approximation, particularly the number of Lanczos iterations, is independent of the number of discrete parameters. Thus, in this example, the number of wave propagation solutions required by the low-rank approximation does not depend on the parameter dimension.

Figures 2.6 and 2.7 show several eigenvectors of the prior preconditioned data misfit Hessian (2.23) (corresponding to several dominant eigenvalues) for Problems I and II. Eigenvectors corresponding to dominant eigenvalues represent the earth modes that are "most observable" from the data, given the configuration of sources and receivers. As can be seen in these figures, the largest eigenvalues produce the smoothest modes, and as the eigenvalues decrease, the associated eigenvectors become more oscillatory, due to the reduced ability to infer smaller length scales from the observations.

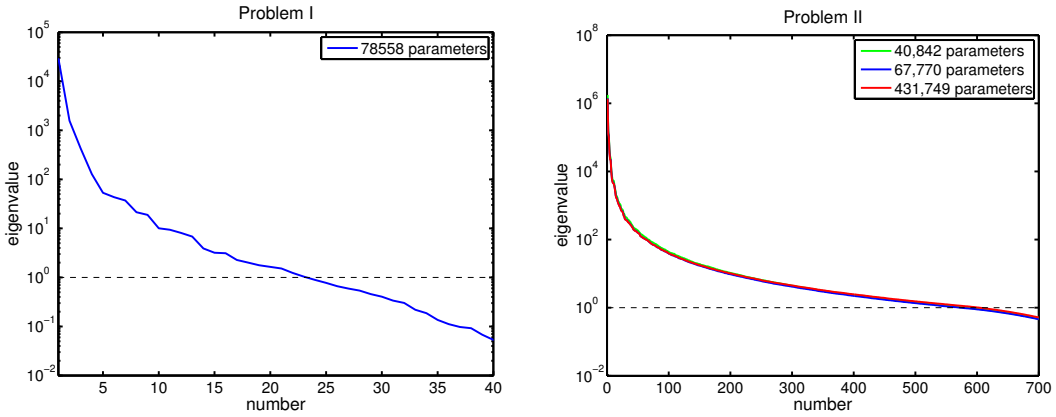

Figure 2.5: Logarithmic plot of the spectrum of prior-preconditioned data misfit Hessian for Problem I (left) and Problem II (right). The computation of the approximate spectrum for Problem I uses a discretization with 78558 wave speed parameters, third-order dG finite elements for the wave propagation solution, and 50 Lanczos iterations. The spectrum for Problem II is computed on different discretizations of the parameter mesh using 900 Lanczos iterations. The eigenvalues for the three discretizations essentially lie on top of each other, which illustrates that the underlying infinite-dimensional statistical inverse problems is properly approximated. The horizontal line $\lambda = 1$ shows the reference value for the truncation of the spectrum of the misfit Hessian. For an accurate approximation of the posterior covariance matrix (i.e., the inverse of the Hessian), eigenvalues that are small compared to 1 can be neglected as discussed in Section 2.5, and in particular as shown in (2.26).

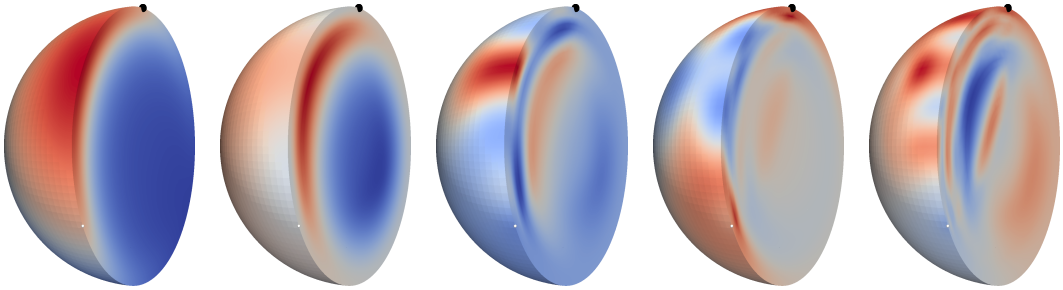

Figure 2.6: Problem I: Eigenvectors of the prior-preconditioned misfit Hessian corresponding to the first (i.e., the largest), the 3rd, the 5th, 8th and 13th eigenvalues (from left to right). The visualization employs a slice through the source and receiver locations.

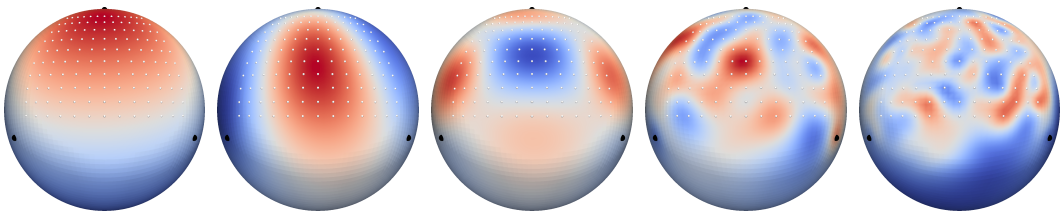

Figure 2.7: Problem II: Eigenvectors of the misfit Hessian corresponding to eigenvalues 1, 5, 20, 100 and 350 respectively. Note that the lower modes are smoothest and become more oscillatory with increasing mode number.

2.6.8 Interpretation of the uncertainty in the solution of the inverse problem

We first study Problem I, i.e., the single source, single receiver problem. Since the data are very sparse, it is expected that we can reconstruct only very limited information from the truth solution; this is reflected in the smoothness of the dominant eigenmodes shown in Figure 2.6. To assess the uncertainty, Figure 2.8 shows prior variance, knowledge gained from the data (i.e., reduction in the variance), and posterior variance, which are computed from (2.29). As discussed in Section 2.5, the posterior is the combination of the prior information and the knowledge gained from the data, so that the posterior uncertainty is decreased relative to the prior uncertainty. That is, the inference has less uncertainty in regions for which the data are more informative. In particular, the region of lowest uncertainty is at the surface half-way between source and receiver, as Figure 2.8 shows. Note that the data are also informative about the core-mantle boundary, where the strong material contrast results in stronger reflected energy back to the surface receivers, allowing greater confidence in the properties of that interface.

Next, we study the results for Problem II. The comparison between the MAP estimate and the ground truth earth model (S20RTS) at different depths is displayed in Figure 2.9. As can be seen, we are able to recover accurately the wave speed in the portion of the Northern hemisphere covered by sources and receivers. We plot the prior and posterior pointwise standard deviations in Figure 2.10. One observes that the uncertainty reduction is greatest along

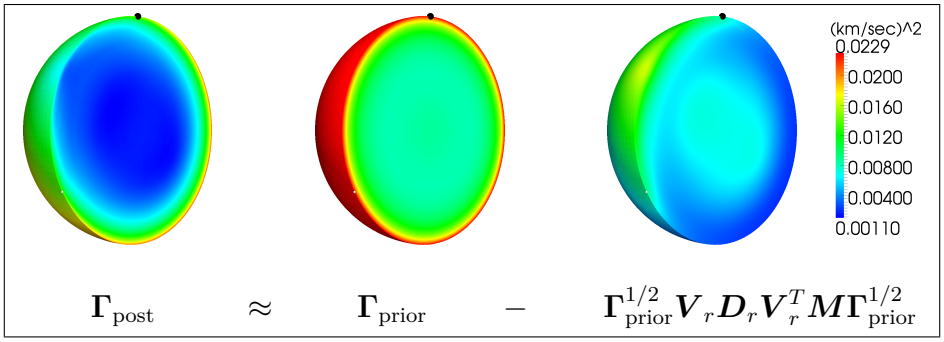

Figure 2.8: Problem I: The left image depicts the pointwise posterior variance field, which is represented as the difference between the original prior variance field (middle), and the reduction in variance due to data (right). The locations of the single source and single receiver is shown by the black and white dot, respectively. Colorscale is common to all three images.

the wave paths between sources and receivers, particularly in the quarter of the Northern hemisphere surface where the receivers are distributed.

In Figure 2.11, we show a comparison between samples from the prior distribution and from the posterior. We observe that in the quarter of the Northern hemisphere where the data are more informative about the medium, we have a higher degree of confidence about the wave speed, which is manifested in the common large scale features across the posterior samples. The fine-scale features (about which the data are least informative) are qualitatively similar to those of the prior distribution, and vary from sample to sample in the posterior. We note that the samples shown here are computed by approximating $\boldsymbol{M}^{-1/2}$ in expression (2.30) using the (diagonal) lumped mass matrix to avoid computing a factorization of $\boldsymbol{M}$. If desired, this mass lumping can be avoided by applying $\boldsymbol{M}^{-1/2}$ to a vector using an iterative scheme based

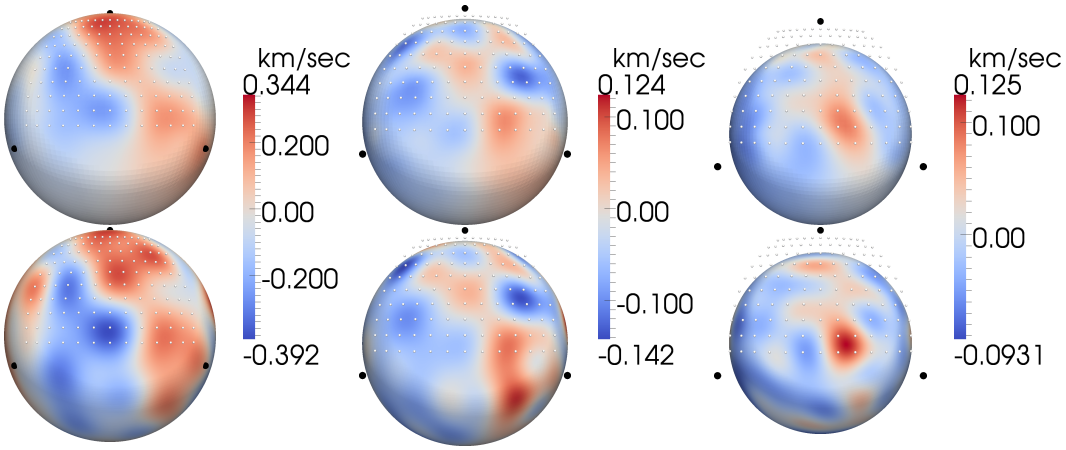

Figure 2.9: Comparison of MAP of posterior pdf (upper row) with the "truth" earth model (lower row) in a depth of $67km$ (left image), $670km$ (middle image) and $1340km$ (right image). The colormap varies with depth, but is held constant between the MAP and "truth" images at each depth.

on polynomial approximations to the matrix function $f(t) = t^{-1/2}$, as in [45].

2.7 Conclusions

A computational framework for estimating the uncertainty in the numerical solution of linearized infinite-dimensional statistical inverse problems is presented. We adopt the Bayesian inference formulation: given observational data and their uncertainty, the governing forward problem and its uncertainty, and a prior probability distribution describing uncertainty in the parameter field, find the posterior probability distribution over the parameter field. The framework, which builds on the infinite-dimensional formulation proposed by Stuart [189], incorporates a number of components aimed at ensuring a convergent discretization of the underlying infinite-dimensional inverse problem. It

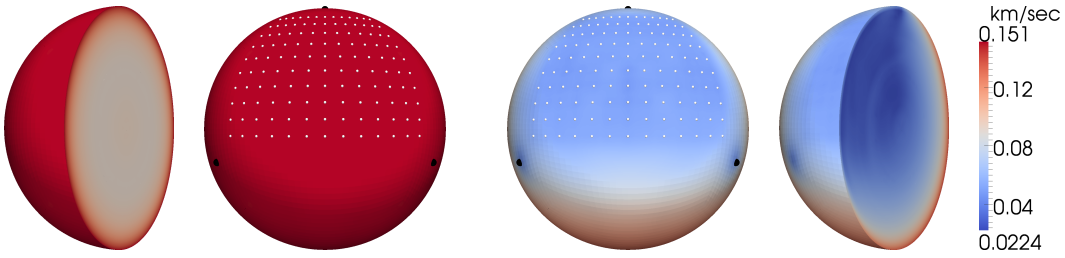

Figure 2.10: The figure compares the pointwise standard deviation for the prior (left) and posterior (right) distributions at a depth of $67km$. The color indicates one standard deviation, and the scale is common to both prior and posterior images. We observe that the most reduction in variance due to data occurs in the region near sources and receivers, whereas the least reduction occurs on the opposite side of the Earth.

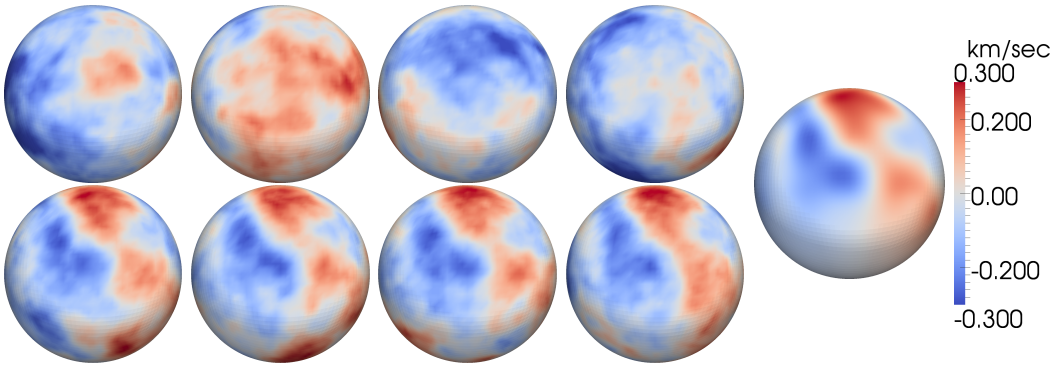

Figure 2.11: Samples from the prior (top row) and posterior (bottom row) distributions. The prior scaling was chosen such that the "ground truth" S20RTS would be a qualitatively reasonable sample from the prior distribution. For comparison purposes, the MAP estimate is shown on the far right.

additionally incorporates algorithms for manipulating the prior, constructing a low rank approximation of the data-informed component of the posterior covariance operator, and exploring the posterior, that together ensure scalability of the entire framework to very high parameter dimensions. Since the data are typically informative about only a low dimensional subspace of the parameter space, the Hessian is sparse with respect to some basis. We have exploited this fact to construct a low rank approximation of the Hessian and its inverse using a parallel matrix-free Lanczos method. Overall, our method requires a dimension-independent number of forward PDE solves to approximate the local covariance. Uncertainty quantification for the linearized inverse problem thus reduces to solving a fixed number of forward and adjoint PDEs (which resemble the original forward problem), independent of the problem dimension. The entire process is thus scalable with respect to the forward problem dimension, uncertain parameter dimension, and observational data dimension. We applied this method to the Bayesian solution of an inverse problem in 3D global seismic wave propagation with up to 430,000 parameters, for which we observe 2–3 orders of magnitude dimension reduction, making UQ for large-scale inverse problems tractable.

2.8 Appendix I: Constructive derivation of square root covariance

In the following, we provide a constructive derivation of $\boldsymbol{L}$ in (2.30) such that it satisfies $\boldsymbol{\Gamma}_{\text{post}} = \boldsymbol{L}\boldsymbol{L}^{\diamond}$. Our goal is to draw posterior Gaussian

random sample with covariance matrix $\boldsymbol{\Gamma}_{\text{post}}$ in $\mathbb{R}^n_{\boldsymbol{M}}$. To accomplish this, a standard approach is first to find a factorization $\boldsymbol{\Gamma}_{\text{post}} = \tilde{\boldsymbol{L}}\tilde{\boldsymbol{L}}^*$, where $\tilde{\boldsymbol{L}}$ is a linear map from $\mathbb{R}^n_{\boldsymbol{M}}$ to $\mathbb{R}^n_{\boldsymbol{M}}$. Then, any random sample from the posterior can be written as

$$\boldsymbol{\nu}^{\text{post}} = \boldsymbol{m}_{\text{MAP}} + \tilde{\boldsymbol{L}}\tilde{\boldsymbol{n}}, \tag{2.38}$$

where $\tilde{\boldsymbol{n}}$ is a Gaussian random sample with zero mean and identity covariance matrix in $\mathbb{R}^n_{\boldsymbol{M}}$. It follows that

$$\tilde{\boldsymbol{n}} = \boldsymbol{M}^{-1/2}\boldsymbol{n},$$

where $\boldsymbol{n}$ is the standard Gaussian random sample with zero mean and identity covariance matrix in $\mathbb{R}^n$, i.e. $\boldsymbol{n} \sim \mathcal{N}(0, \boldsymbol{I})$, and $\boldsymbol{M}^{-1/2}$ a linear map from $\mathbb{R}^n$ to $\mathbb{R}^n_{\boldsymbol{M}}$.

Therefore, what remains to be done is to construct $\tilde{\boldsymbol{L}}$. To begin the construction, we rewrite (2.28) as

$$\boldsymbol{\Gamma}_{\text{post}} \approx \boldsymbol{\Gamma}^{1/2}_{\text{prior}} \underbrace{\left(\boldsymbol{I} - \boldsymbol{V}_r\boldsymbol{D}_r\boldsymbol{V}_r^\diamond\right)}_{\boldsymbol{B}} \boldsymbol{\Gamma}^{1/2}_{\text{prior}}.$$

The simple structure of $\boldsymbol{B}$ allows us to write its spectral decomposition as

$$\boldsymbol{B} = \sum_{i=1}^{n} \boldsymbol{v}_i\boldsymbol{v}_i{}^\diamond - \sum_{i=1}^{r} \frac{\lambda_i}{\lambda_i + 1}\boldsymbol{v}_i\boldsymbol{v}_i{}^\diamond = \sum_{i=1}^{r} \frac{1}{\lambda_i + 1}\boldsymbol{v}_i\boldsymbol{v}_i{}^\diamond + \sum_{i=r+1}^{n} \boldsymbol{v}_i\boldsymbol{v}_i{}^\diamond,$$

which, together with the standard definition of the square root of postive self-adjoint operators [9], immediately gives

$$\boldsymbol{B}^{1/2} = \sum_{i=1}^{r} \frac{1}{\sqrt{\lambda_i + 1}}\boldsymbol{v}_i\boldsymbol{v}_i{}^\diamond + \sum_{i=r+1}^{n} \boldsymbol{v}_i\boldsymbol{v}_i{}^\diamond = \sum_{i=1}^{r} \left(\frac{1}{\sqrt{\lambda_i + 1}} - 1\right)\boldsymbol{v}_i\boldsymbol{v}_i{}^\diamond + \sum_{i=1}^{n} \boldsymbol{v}_i\boldsymbol{v}_i{}^\diamond = \boldsymbol{V}_r\boldsymbol{P}_r\boldsymbol{V}_r^\diamond + \boldsymbol{I},$$

where $\boldsymbol{P}_r = \operatorname{diag}\left(1/\sqrt{\lambda_1+1}-1,\ldots,1/\sqrt{\lambda_r+1}-1\right) \in \mathbb{R}^{r\times r}$, and $\boldsymbol{B}^{1/2}$ is self-adjoint in $\mathbb{R}^n_{\boldsymbol{M}}$, namely, $\left(\boldsymbol{B}^{1/2}\right)^* = \boldsymbol{B}^{1/2}$. Now, we define

$$\tilde{\boldsymbol{L}} = \boldsymbol{\Gamma}_{\text{prior}}^{1/2}\boldsymbol{B}^{1/2},$$

which, by construction, is the desired matrix owing to the following trivial identity

$$\tilde{\boldsymbol{L}}\tilde{\boldsymbol{L}}^* = \boldsymbol{\Gamma}_{\text{prior}}^{1/2}\boldsymbol{B}^{1/2}\left(\boldsymbol{B}^{1/2}\right)^*\left(\boldsymbol{\Gamma}_{\text{prior}}^{1/2}\right)^* = \boldsymbol{\Gamma}_{\text{prior}}^{1/2}\boldsymbol{B}\boldsymbol{\Gamma}_{\text{prior}}^{1/2} = \boldsymbol{\Gamma}_{\text{post}},$$

where we have used the self-adjointness of $\boldsymbol{\Gamma}_{\text{prior}}^{1/2}$ and $\boldsymbol{B}^{1/2}$ in $\mathbb{R}^n_{\boldsymbol{M}}$.

Finally, we can rewrite (2.38) in terms of $\boldsymbol{n}$ and $\boldsymbol{L} = \tilde{\boldsymbol{L}}\boldsymbol{M}^{-1/2}$, a linear map from $\mathbb{R}^n$ to $\mathbb{R}^n_{\boldsymbol{M}}$, as

$$\boldsymbol{\nu}^{\text{post}} = \boldsymbol{m}_{\text{MAP}} + \boldsymbol{L}\boldsymbol{n},$$

where $\boldsymbol{L}$ satisfies the following desired identity

$$\boldsymbol{L}\boldsymbol{L}^{\diamond} = \tilde{\boldsymbol{L}}\boldsymbol{M}^{-1/2}\left(\boldsymbol{M}^{-1/2}\right)^{\diamond}\tilde{\boldsymbol{L}}^* = \tilde{\boldsymbol{L}}\boldsymbol{M}^{-1/2}\boldsymbol{M}^{-1/2}\boldsymbol{M}\tilde{\boldsymbol{L}}^* = \tilde{\boldsymbol{L}}\tilde{\boldsymbol{L}}^* = \boldsymbol{\Gamma}_{\text{post}}.$$

2.9 Appendix II: Framework extensions for mild nonlinearity

In this section, we describe strategies for characterization of the true posterior distribution when the nonlinearity of the parameter-to-observable map $\boldsymbol{f}(\boldsymbol{m})$ is mild over the support of the posterior density π_{post}. To do so use the Gaussian density with mean at the MAP estimate $\boldsymbol{m}_{\text{MAP}}$, and covariance $\boldsymbol{\Gamma}_{\text{post}}$ from equation (2.28) as an approximation $\tilde{\pi}_{\text{post}}$ to the true posterior density directly. Finally, we characterize the true posterior density using N samples $\boldsymbol{m}_k$ for $k = 0, \ldots, N-1$, from the approximation $\tilde{\pi}_{\text{post}}$ and evaluating $\pi_{\text{post}}(\boldsymbol{m}_k)$ for each sample. Provided that the assumption of mild nonlinearity in $\boldsymbol{f}$ is valid, this is sufficient to solve the full statistical inverse problem.

In the following subsections, we describe the background for importance sampling and the independence sampler MCMC methods. We next discuss the additional computational challenges that arise for large-scale problems in this context. Finally, we apply both methods to the example problem in global seismic inversion.

2.9.1 Sampling Methods

We first compute several sample realizations from the approximate posterior density using equation (2.31), and compute the value w_k for each sample,

$$w_k := \frac{\pi_{\text{post}}(\boldsymbol{m}_k)}{\tilde{\pi}_{\text{post}}(\boldsymbol{m}_k)}. \tag{2.39}$$

If indeed $\tilde{\pi}_{\text{post}}$ is a good approxmation to π_{post} (up to an unknown constant), then the values of w_k will not be too disparate, and the collection of samples

$\boldsymbol{m}_k$ and corresponding values w_k will be sufficient to characterize π_{post}.

It is notable that because the samples $\boldsymbol{m}_k$ are independent, they may be generated and the corresponding value w_k evaluated offline and in parallel, even if they are subsequently to be interpreted in the context of MCMC. This can be of practical value in high performance computing as this helps them to be robust to failure and restarting if a computation is aborted for any reason. The disadvantage to this approach is that the posterior approximation is required to be quite good for this approach to be successful, and if no such approximation is available then more sophisticated approaches such as those applied in [166] are recommended.

2.9.1.1 Importance Sampling

In the importance sampling framework, each sample $\boldsymbol{m}_k$ is assigned a sample weight corresponding to the value w_k, and desired statistics of the posterior distribution are then computed using weighted averages of these samples. To illustrate this approach, consider a quantity of interest $\phi(\boldsymbol{m})$ that we desire to estimate from the posterior distribution. We calculate,

$$\mathbb{E}[\phi(\boldsymbol{m})] = \int \phi(\boldsymbol{m})\pi_{\text{post}}(\boldsymbol{m})d\boldsymbol{m} \tag{2.40}$$

$$= \int \phi(\boldsymbol{m})\frac{\pi_{\text{post}}(\boldsymbol{m})}{\tilde{\pi}_{\text{post}}(\boldsymbol{m})}\tilde{\pi}_{\text{post}}(\boldsymbol{m})d\boldsymbol{m} \tag{2.41}$$

$$= \int \phi(\boldsymbol{m})w(\boldsymbol{m})\tilde{\pi}_{\text{post}}(\boldsymbol{m})d\boldsymbol{m} \tag{2.42}$$

$$\approx \frac{1}{N}\sum_{k=0}^{N-1}\phi(\boldsymbol{m}_k)w_k. \tag{2.43}$$

Some caution is advisable to ensure that $\phi(\boldsymbol{m})w(\boldsymbol{m})$ has finite variance under the approximate distribution to ensure the Monte Carlo-type estimate in (2.43) converges under the law of large numbers. In particular, this can fail to be the case in practice if the approximate posterior has shorter tails than the true posterior. This could influence the construction of the approximate posterior density somewhat, since it is crucial that we overestimate the variance in $\tilde{\pi}_{\text{post}}$ when we are not certain of the correct values.

In many cases where the true posterior is nearly Gaussian, we anticipate that the majority of the error in our approximation to the posterior results from the truncation that occurs in our low-rank construction of the prior-preconditioned data misfit Hessian. Because this enters into the final expression for the posterior covariance in equation (2.27) as a negative definite update to the prior covariance, we expect this truncation to strictly overestimate the true posterior variance in the subspace associated with the truncated modes.

Further analytical results can be found in the literature in the asymptotic limit of small observation noise. In particular, see [91] for a detailed analysis in the context of implicit sampling.

2.9.1.2 Markov Chain Monte Carlo Sampling of the posterior

Markov chain Monte Carlo provides a general means to sample from the posterior distribution π_{post} without appealing to quadrature in high dimensions. Instead, a *proposal* distribution which can be sampled easily is

Algorithm 2 Metropolis-Hastings MCMC algorithm to sample the pdf π_{post}

Choose initial parameters $\boldsymbol{m}_0$
Compute $\pi_{\text{post}}(\boldsymbol{m}_0)$
for $k = 0, \ldots, N-1$ **do**
  Draw $\boldsymbol{m}_{\text{proposal}}$ from the proposal density $q(\boldsymbol{m}_k, \boldsymbol{m}_{\text{proposal}}) = \tilde{\pi}_{\text{post}}(\boldsymbol{m}_{\text{proposal}})$
  Compute $\pi_{\text{post}}(\boldsymbol{m}_{\text{proposal}})$
  Compute $\alpha_k = \min\left\{1, \frac{\pi_{\text{post}}(\boldsymbol{m}_{\text{proposal}})\tilde{\pi}_{\text{post}}(\boldsymbol{m}_k)}{\pi_{\text{post}}(\boldsymbol{m}_k)\tilde{\pi}_{\text{post}}(\boldsymbol{m}_{\text{proposal}})}\right\} = \min\left\{1, \frac{w(\boldsymbol{m}_{\text{proposal}})}{w(\boldsymbol{m}_k)}\right\}$
  Draw $u \sim \mathcal{U}([0,1])$
  if $u < \alpha_k$ **then**
    Accept: Set $\boldsymbol{m}_{k+1} = \boldsymbol{m}_{\text{proposal}}$
  else
    Reject: Set $\boldsymbol{m}_{k+1} = \boldsymbol{m}_k$ (i.e., this sample is repeated in the resulting MCMC chain)
  end if
end for

employed. We use our approximation $\tilde{\pi}_{\text{post}}$ as this proposal distribution, and each proposed sample $\boldsymbol{m}_k$ is subjected to the Metropolis-Hastings accept/reject framework with an acceptance probability that depends only on the previously computed values $w(\boldsymbol{m}_k)$. Under relatively mild conditions, it can be shown that the resulting sample chain converges to the desired posterior distribution. Additionally, there are in principal no restrictions on the quantities of interest that we wish to estimate as there were with importance sampling.

The Metropolis-Hastings algorithm [104,149] is outlined in algorithm 2. Different MCMC methods are distinguished by the choice of proposal density $q(\boldsymbol{m}_k, \boldsymbol{m}_{\text{proposal}})$. The proposal here is chosen to be our approximate posterior density,

$$q(\boldsymbol{m}_k, \boldsymbol{m}_{\text{proposal}}) = \tilde{\pi}_{\text{post}}(\boldsymbol{m}_{\text{proposal}}) \approx \pi_{\text{post}}(\boldsymbol{m}_{\text{proposal}}), \tag{2.44}$$

and therefore the acceptance probability is

$$\alpha_k = \min\left(1, \left\{\frac{w(\boldsymbol{m}_{\text{proposal}})}{w(\boldsymbol{m}_k)}\right\}\right) \approx 1, \tag{2.45}$$

and we have nearly 100% acceptance rate if the values $w(\boldsymbol{m}_k)$ do not vary much. Because the proposal in (2.44) is independent of the current MCMC state $\boldsymbol{m}_k$, the resulting method is termed an *independence sampler*.

2.9.2 Computational considerations

Most of the necessary computational tools for these sampling methods are the same as those required for the computational framework discussed in this chapter. Here, we address only the additional tools that are required for sample generation from the approximate posterior, and evaluation of the true posterior density at a given sample point.

To generate samples from the approximate posterior distribution, our algorithms require the application of $\boldsymbol{M}^{-1/2}$, the inverse square root of the mass matrix, or more generally a matrix $\boldsymbol{L}$ such that $\boldsymbol{L}\boldsymbol{L}^T = \boldsymbol{M}^{-1}$. Since the explicit computation of the matrix square root is expensive in high dimensions, we utilize an iterative algorithm to compute the application of the inverse matrix square root to vectors [45]. This method relies on the approximation of the square root via orthogonal polynomials on an interval containing the spectrum of the matrix. Convergence is most quickly obtained when this interval is as small as possible, so we choose to first symmetrically precondition $\boldsymbol{M}$ with the lumped mass matrix $\boldsymbol{M}_l$. The eigenvalues of the resulting

matrix $\tilde{\boldsymbol{M}} = \boldsymbol{M}_l^{-1/2}\boldsymbol{M}\boldsymbol{M}_l^{-1/2}$ are clustered around 1, resulting in fast convergence when applying it to vectors as discussed in [45]. Finally, we can take $\boldsymbol{L} := \boldsymbol{M}_l^{-1/2}\tilde{\boldsymbol{M}}^{-1/2}$, which satisfies the original requirement for $\boldsymbol{L}$ and is computationally efficient. All other aspects of sampling are straightforward and described in the main text of this chapter.

It remains to evaluate the value of w_k at each sample point. The only potential concern here is a numerical one, as direct evaluation of $\pi_{\text{post}}(\boldsymbol{m}_k)$ using equation (2.19) is very simple. Consider the computation of w_k,

$$\log w_k \tag{2.46a}$$

$$= \log\left(\frac{\pi_{\text{post}}(\boldsymbol{m}_k)}{\tilde{\pi}_{\text{post}}(\boldsymbol{m}_k)}\right) \tag{2.46b}$$

$$= \log\left(\pi_{\text{post}}(\boldsymbol{m}_k)\right) - \log\left(\tilde{\pi}_{\text{post}}(\boldsymbol{m}_k)\right) \tag{2.46c}$$

$$= -\frac{1}{2}\|\boldsymbol{m}_k - \boldsymbol{m}_{\text{MAP}}\|^2_{\boldsymbol{\Gamma}^{-1}_{\text{post}}} + \frac{1}{2}\|\boldsymbol{f}(\boldsymbol{m}_k) - \boldsymbol{y}^{\text{obs}}\|^2_{\boldsymbol{\Gamma}^{-1}_{\text{noise}}} + \frac{1}{2}\|\boldsymbol{m}_k - \boldsymbol{m}_0\|^2_{\boldsymbol{\Gamma}^{-1}_{\text{prior}}} \tag{2.46d}$$

$$= -\left(\frac{1}{2}\|\boldsymbol{m}_k - \boldsymbol{m}_0\|^2_{\boldsymbol{\Gamma}^{-1}_{\text{prior}}} - \langle \boldsymbol{m}_k - \boldsymbol{m}_{\text{MAP}}, \boldsymbol{\Gamma}^{1/2}_{\text{prior}}\boldsymbol{V}_r\boldsymbol{D}_r\boldsymbol{V}_r^{\diamond}\boldsymbol{\Gamma}^{1/2}_{\text{prior}}(\boldsymbol{m}_k - \boldsymbol{m}_{\text{MAP}})\rangle\right)$$
$$+ \frac{1}{2}\|\boldsymbol{f}(\boldsymbol{m}_k) - \boldsymbol{y}^{\text{obs}}\|^2_{\boldsymbol{\Gamma}^{-1}_{\text{noise}}} + \frac{1}{2}\|\boldsymbol{m}_k - \boldsymbol{m}_0\|^2_{\boldsymbol{\Gamma}^{-1}_{\text{prior}}} \tag{2.46e}$$

$$= \langle \boldsymbol{m}_k - \boldsymbol{m}_{\text{MAP}}, \boldsymbol{\Gamma}^{1/2}_{\text{prior}}\boldsymbol{V}_r\boldsymbol{D}_r\boldsymbol{V}_r^{\diamond}\boldsymbol{\Gamma}^{1/2}_{\text{prior}}(\boldsymbol{m}_k - \boldsymbol{m}_{\text{MAP}})\rangle + \frac{1}{2}\|\boldsymbol{f}(\boldsymbol{m}_k) - \boldsymbol{y}^{\text{obs}}\|^2_{\boldsymbol{\Gamma}^{-1}_{\text{noise}}}. \tag{2.46f}$$

The simplification after (2.46e) is done to avoid cancellation error from the log prior terms, $\frac{1}{2}\|\boldsymbol{m}_k - \boldsymbol{m}_0\|^2_{\boldsymbol{\Gamma}^{-1}_{\text{prior}}}$. These terms grow (in expected value) with the discretized parameter dimension n as $O(n)$, while $\log w_k$ is independent of n, i.e., $O(1)$. At sufficiently large n, this cancellation error will overwhelm the final result.

That $\log w_k$ is independent of n is not necessarily obvious. However, under the assumption that π_{post} is well approximated by $\tilde{\pi}_{\text{post}}$, we expect the values of $\log w_k$ to be near in value to each other, and in particular near in value to the value we would compute at the MAP estimate,

$$\log(w(\boldsymbol{m}_{\text{MAP}})) = \frac{1}{2}\|\boldsymbol{f}(\boldsymbol{m}_k) - \boldsymbol{y}^{\text{obs}}\|^2_{\boldsymbol{\Gamma}^{-1}_{\text{noise}}} = \pi_{\text{like}}(\boldsymbol{y}^{\text{obs}}|\boldsymbol{m}_{\text{MAP}}). \tag{2.47}$$

Because the likelihood function is a well-defined infinite-dimensional quantity and we have been careful to be sure that our discretization converges to the desired infinite-dimensional solution with increasing n (i.e., refinement of the parameter mesh), $\log(w(\boldsymbol{m}_{\text{MAP}}))$ converges to a value independent of n. From here, it is reasonable to expect also that the values w_k also do not grow with increasing n.

2.9.3 Numerical results

In this section, we discuss the numerical results for the nonlinear statistical inverse problem in both the independence sampler MCMC and the importance sampling contexts. Because both methods discussed here only require independent draws from the proposal distribution, we are free to generate a large number of samples in advance (and in parallel, if desired), and for each sample $\boldsymbol{m}_k$, evaluate the weights w_k. We generated 15,587 samples from the proposal distribution and evaluate the weights using 2,048 cores on TACC's Lonestar system, in a total of approximately 96 hours of runtime. The dominant computational cost is evaluation of the true posterior density, which requires a forward wave simulation. All further computations for both

independence sampler MCMC and importance sampling are performed easily on a workstation.

Because we are evaluating both methods using the same dataset, it is likely uninformative and potentially misleading to compare performance characteristics of these methods against each other. The reason for presenting results based on both approaches is to demonstrate applicability of the methods in different contexts, and to compute convergence diagnostics with which the reader may be most familiar.

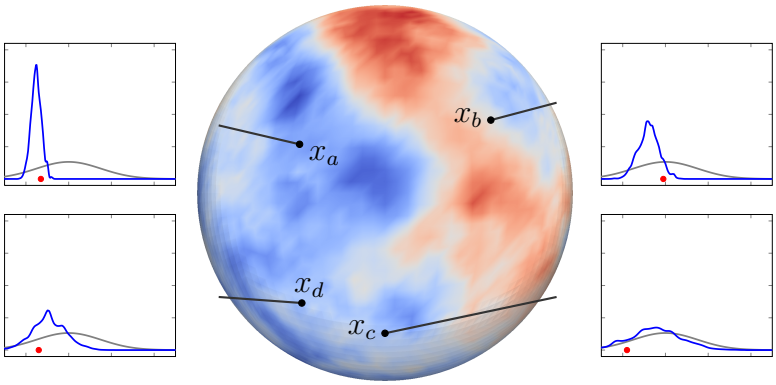

Figure 2.12: Illustrative sample from the posterior distribution generated by the independence sampler MCMC method. On the surface at each of the black dots, the pointwise 1D marginal distribution for both the estimated posterior distribution (blue) and the Gaussian prior distribution (grey) are shown. The red dot on the horizontal axis of each plot represents the parameter value of the illustrated sample at the given location. We observe that far from the measurement locations in the southern hemisphere, the 1D marginal distributions have high variance, and do not differ significantly from those of the prior distribution. In the northern hemisphere, the data are much more informative about the parameter values in these locations, and the resulting 1D marginal posterior distributions are sharply peaked compared to the prior distribution.

2.9.4 MCMC Results

In this section, we discuss the numerical results of an independence sampler MCMC method using the precomputed samples and weights discussed above. Of the initial 15,587 samples, 4,351 were accepted, for an overall acceptance rate of 28%. To assess convergence of the MCMC chain, we can visually inspect the parameter chain at a number of different indices to see how well the chain is mixing across the distribution. This is shown in figure 2.13.

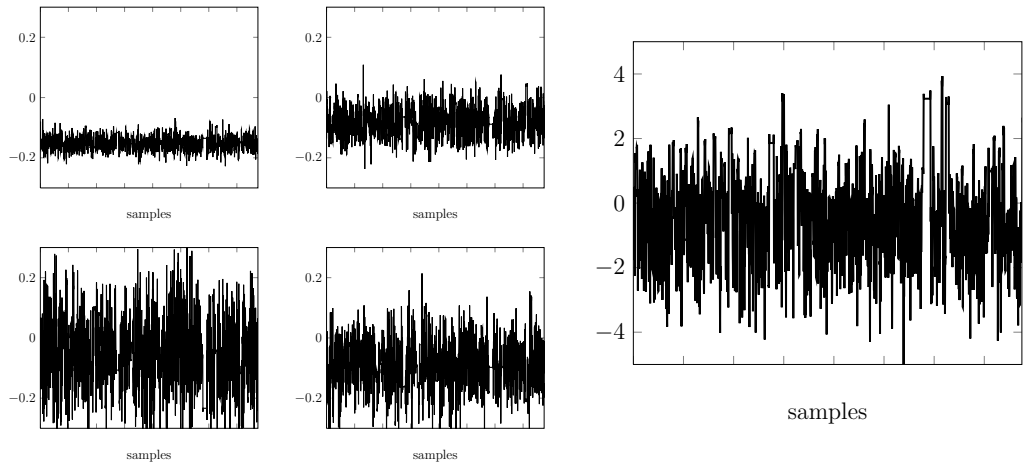

Figure 2.13: Traces for various surface coordinate parameters for MCMC chain. Rapid mixing of the chains across the distribution indicate good convergence. In the left four panels, we show trace plots for the points $x_a, \ldots, x_d$ shown in figure 2.12. In the right plot, we show the trace of the MCMC weight used in the Metropolis-Hastings accept/reject step, with higher values indicating higher probability of acceptance.

A more quantitative assessment of this chain mixing is available using the integrated autocorrelation time τ. In traditional Monte Carlo methods, averaging over N i.i.d. samples from the true posterior distribution would re-

sult in a variance reduction of $\frac{1}{N}$ for the given quantity of interest. In MCMC, samples become correlated by the accept/reject framework of Metropolis Hastings, and will not achieve this $\frac{1}{N}$ reduction in variance. The integrated autocorrelation time provides an estimate of the number of MCMC samples for a particular problem that would be required to achieve the same variance reduction as classical Monte Carlo. The integrated autocorrelation time is computed as

$$\tau = 1 + 2\sum_{s=1}^{\infty} \rho(s), \tag{2.48}$$

where $\rho(s)$ is the usual autocorrelation function for a lag s. In practice for finite length sample chains, $\rho(s)$ is a noisy function, and a tradeoff must be made when estimating τ; we must retain enough terms for the summation to converge, but every additional term in the summation increases the variance in our estimate. Visual inspection of $\rho(s)$ for this problem indicates that 500 terms in the summation is sufficient for this sum to converge and then be overwhelmed by noise, and we report the *maximum* partial sum for τ obtained over this interval. As such, we expect to overestimate the true value of τ in most cases.

Finally, we can compute the effective sample size,

$$\text{ESS} = \frac{N}{\tau}, \tag{2.49}$$

which allows for direct comparison between our MCMC and the number of equivalent independent samples that would produce the same reduction in variance. The effective sample size depends on the desired quantity of interest

	x_a	x_b	x_c	x_d	w
IAT	26.7	20.4	31.1	19.5	63.5
ESS	584	764	501	799	245

Table 2.1: Integrated autocorrelation time (IAT) and the corresponding effective sample size (ESS) for our MCMC estimates of the medium using 15,587 samples. We report statistics for estimators at four points x_a, x_b, x_c and x_d shown in Figure 2.12, as well as the estimator for the weight used in the accept/reject step for Metropolis-Hastings.

as different quantities of interest for the same MCMC chain can mix quite differently. We present integrated autocorrelation time and effective sample size for several quantities of interest in table 2.1.

2.9.5 Importance sampling results

Next, we consider the same set of samples in an importance sampling framework, where each sample is given a weight reflecting its relative contribution to the posterior distribution.

Because both the true posterior distribution π_{post} and the approximate distribution $\tilde{\pi}_{\text{post}}$ are each known only up to a constant value, it is necessary to self-normalize the importance weights. Here, we choose to normalize the weights such that $\sum_{k=0}^{N-1} w_k = N$. This way, if $\tilde{\pi}_{\text{post}}$ is in fact equivalent to π_{post} (up to the unknown constant), then all renormalized weights will be 1 and the method is equivalent to using Monte Carlo to sample the true posterior distribution. To the extent that these weights are not uniform, this can provide an estimate of how well our samples approximate the true distribution. Figure

2.14 displays the sorted weights compared to the constant value 1.

In general, importance sampling is effective for our purposes when the computed weights are not too disparate. Intuitively, this can be understood by looking at the two extreme cases. If all samples had equal weights, then (after normalization) all weights would be effectively equal to 1, and this method is equivalent to perfect sampling from π_{post}. On the other hand, if the computed weights were so disparate that one value of w_k dominated the sum of all other sample weights, then we would have only a single effective sample in our set, and we would be unwise to trust it anyway. For the current sample set, the ratio between the largest individual sample weight and the total weight is 0.0116, so even the most significant samples only account for about 1% of the available information.

This concept is made quantitative in the coefficient of variation of the weights, $\text{cv}(w)$, and can be used to compute an effective sample size[5] for our problem, as in [158]. The coefficient of variation and effective sample size are

[5]The ESS value computed here corresponds only to the final column of the MCMC results in table 2.1. Corresponding expressions for specific quantities of interest (i.e., the first four columns of table 2.1) can also be found in [158]. At the time of publication, these QoI-specific results for the importance sampling approach are not available.

computed as

$$\mathrm{cv}(w) = \frac{1}{\bar{w}} \left(\frac{1}{N-1} \sum_{k=0}^{N-1} (w_k - \bar{w})^2 \right)^{1/2}, \tag{2.50}$$

$$\bar{w} = \frac{1}{N} \sum_{k=0}^{N-1} w_k, \tag{2.51}$$

$$\mathrm{ESS} = \frac{N}{1 + \mathrm{cv}(w)^2}. \tag{2.52}$$

For this problem, we have $\mathrm{cv}(w) = 3.22$ and $\mathrm{ESS} = 1370$.

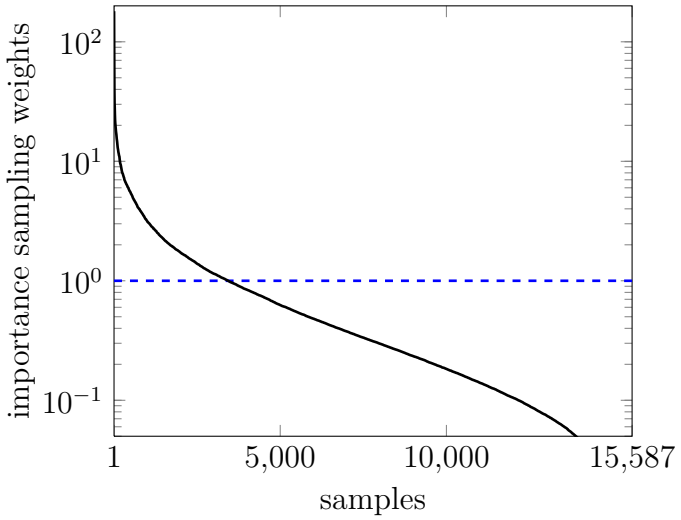

Figure 2.14: The black curve in this plot shows the sorted and self-normalized importance sampling weights for this problem. The blue horizontal line represents a fixed weight of 1.0 for all samples, which would be the optimal setting for Monte Carlo. Deviation from this value gives some indication about the quality of our importance sampling results.

Chapter 3

Extreme-Scale UQ for Bayesian Inverse Problems Governed by PDEs

The content of this chapter is based on an existing publication[1] [27], which is joint work with Tan Bui-Thanh, Carsten Burstedde, Georg Stadler, Lucas C. Wilcox, and my advisor Omar Ghattas. Georg contributed most of the effort in setting up the deterministic wave propagation and observation operators for the numerical experiments in this chapter, and managing the large-scale computations and scaling experiments. Georg and I collaborated on the implementation of the algorithms for statistical inversion, and finally interpretation and visualization of the results. All authors had significant contribution to the remaining content of this chapter.

Abstract

Quantifying uncertainties in large-scale simulations has emerged as the central challenge facing CS&E. When the simulations require supercomputers,

[1] © 2012 IEEE. Reprinted, with permission, from T. Bui-Thanh, C. Burstedde, O. Ghattas, J. Martin, G. Stadler, and L. C. Wilcox. Extreme-scale UQ for Bayesian inverse problems governed by PDEs. In *SC12: Proceedings of the International Conference for High Performance Computing, Networking, Storage and Analysis*, November 2012. http://dx.doi.org/10.1109/SC.2012.56

and uncertain parameter dimensions are large, conventional UQ methods fail. Here we address uncertainty quantification for large-scale inverse problems in a Bayesian inference framework: given data and model uncertainties, find the pdf describing parameter uncertainties. To overcome the curse of dimensionality of conventional methods, we exploit the fact that the data are typically informative about low-dimensional manifolds of parameter space to construct low rank approximations of the covariance matrix of the posterior pdf via a matrix-free randomized method. We obtain a method that scales independently of the forward problem dimension, the uncertain parameter dimension, the data dimension, and the number of cores. We apply the method to the Bayesian solution of an inverse problem in 3D global seismic wave propagation with over one million uncertain earth model parameters, 630 million wave propagation unknowns, on up to 262K cores, for which we obtain a factor of over 2000 reduction in problem dimension. This makes UQ tractable for the inverse problem.

3.1 Introduction

Perhaps the central challenge facing the field of computational science and engineering today is: *how do we quantify uncertainties in the predictions of our large-scale simulations, given limitations in observational data, computational resources, and our understanding of physical processes* [155]. For many societal grand challenges, the "single point" deterministic predictions delivered by most contemporary large-scale simulations of complex systems are

just a first step: to be of value for decision-making (design, control, allocation of resources, policy-making, etc.), they must be accompanied by the degree of confidence we have in the predictions. Examples of problems for which large-scale simulations are playing an increasingly important role for decision-making include: mitigation of global climate change, natural hazard forecasts; siting of nuclear waste repositories, monitoring of subsurface contaminants, control of carbon sequestration processes, management of the nuclear fuel cycle, design of new nano-structured materials and energy storage systems, and patient-specific planning of surgical procedures, to name a few.

Unfortunately, when the simulations (here assumed without loss of generality to comprise PDEs) are expensive, and the uncertain parameter dimension is large (or even just moderate), conventional uncertainty quantification methods fail dramatically. Here we address uncertainty quantification (UQ) in large-scale inverse problems governed by PDEs. This is the crucial step in UQ: before we can propagate parameter uncertainties forward through a model, we must first infer them from observational data and from the (PDE) model that maps parameters to observables; i.e., we must solve the inverse problem. We adopt the Bayesian inference framework [117, 192]: given observational data and their uncertainty, the governing forward PDEs and their uncertainty, and a prior probability distribution describing prior uncertainty in the parameters, find the posterior probability distribution over the parameters, which is seen as the solution of the inverse problem. The grand challenge in solving statistical inverse problems is in computing statistics of the posterior probability density

function (pdf), which is a surface in high dimensions. This is notoriously challenging for statistical inverse problems governed by expensive forward models (as in our target case of global seismic wave propagation) and high-dimensional parameter spaces (as in our case of inferring a heterogeneous parameter field). The difficulty stems from the fact that evaluation of the probability of each point in parameter space requires solution of the forward problem (which may tax contemporary supercomputers), and many such evaluations (millions or more) are required to adequately sample the posterior density in high dimensions by conventional Markov-chain Monte Carlo (MCMC) methods. Thus, UQ for the large-scale inverse problems becomes intractable.

The approach we take is based on a linearization of the parameter-to-observable map, which yields a local Gaussian approximation of the posterior. The mean and covariance of this Gaussian can be found from an appropriately weighted regularized nonlinear least squares optimization problem, which is known as the *maximum a posteriori* (MAP) point. The solution of this optimization problem provides the mean, and the inverse of the Hessian matrix of the least squares function (evaluated at the MAP point) gives the covariance matrix. Unfortunately, the most efficient algorithms available for direct computation of the (nominally dense) Hessian are prohibitive, requiring as many forward PDE-like solves as there are uncertain parameters, which can number in the millions or more when the parameter represents a field (e.g, initial condition, heterogeneous material coefficient, source term).

The key insight to overcoming this barrier is that the data are typically

informative about a low dimensional manifold of the parameter space [30]—that is, the Hessian of the data-misfit term in the least squares function is sparse with respect to some basis. We exploit this fact to construct a low rank approximation of the data-misfit Hessian and the resulting posterior covariance matrix using a parallel, matrix-free randomized algorithm, which requires a *dimension-independent* number of forward PDE solves and associated adjoint PDE solves (the latter resemble the forward PDEs in reverse time). UQ thus reduces to solving a fixed (and often small, relative to the parameter dimension) number of PDEs. When scalable solvers are available for the forward PDEs, *the entire process of quantifying uncertainties in the solution of the inverse problem is scalable with respect to PDE state variable dimension, uncertain parameter dimension, observational data dimension, and number of processor cores.* We apply this method to the Bayesian solution of an inverse problem in 3D global seismic wave propagation with 1.067 million parameters and 630 million wave propagation spatial unknowns over 2400 time steps, on up to 262,144 Jaguar cores. The example demonstrates independence of parameter dimension and a factor of over 2000 reduction in problem dimension. This UQ computation is orders of magnitude larger than any attempted before on a large-scale forward problem.

We recently presented a finite-dimensional version of our method (in which Lanczos iterations are used to build the low rank approximation of the Hessian) and applied it to a 1D inverse problem in moderate dimensions [143]. We have also recently described the extension to infinite-dimensional inverse

problems (so-called because the inversion parameters represent a field) in the framework of [189], in which we discuss mathematically subtle yet critical issues related to the proper choice of prior and to discretizations that assure convergence to the correct infinite-dimensional quantities [33]. In this, our Bell Prize submission in the Scalable Algorithms category, we extend the method to extreme-scale Bayesian inverse problems, employing a randomized parallel matrix-free low rank approximation method, instead of Lanczos. The randomized method yields a low rank approximation with controllably high probability, and is asynchronous, more robust, more fault tolerant, and provides better cache performance. In the following sections, we provide an overview of the Bayesian formulation of inverse problems (§3.2), describe how the mean and covariance of the posterior pdf can be approximated from the solution of a regularized weighted nonlinear least-squares problem (§3.3 and §3.4), present our algorithm for parallel low rank-based covariance approximation (§3.5), assert the scalability of the overall UQ method (§3.6), apply our method to the Bayesian solution of a very large scale inverse problem in 3D global seismic wave propagation (§3.7), and draw conclusions (§3.8).

3.2 Bayesian Formulation of Inverse Problems

In the Bayesian approach, we state the inverse problem as a problem of *statistical inference* over the space of uncertain parameters, which are to be inferred from the data and a PDE model. The resulting solution to the statistical inverse problem is a posterior distribution that assigns to any candidate

set of parameter fields our belief (expressed as a probability) that a member of this candidate set is the "true" parameter field that gave rise to the observed data. When discretized, this problem of infinite dimensional inference gives rise naturally to a large scale problem of inference over the discrete parameter space $\boldsymbol{x} \in \mathbb{R}^n$, corresponding to degrees of freedom in the parameter field mesh. While the presentation in this paper is limited to the finite dimensional approximation to the infinite dimensional measure, the discretization process is performed rigorously following [33, 189], and the numerical evidence indicates that we converge to the correct infinite dimensional distribution.

The posterior probability distribution combines the prior pdf $\pi_{\text{prior}}(\boldsymbol{x})$ over the parameter space, which encodes any knowledge or assumptions about the parameter space that we may wish to impose before the data are considered, with a likelihood pdf $\pi_{\text{like}}(\boldsymbol{y}_{\text{obs}}|\boldsymbol{x})$, which explicitly represents the probability that a given set of parameters $\boldsymbol{x}$ might give rise to the observed data $\boldsymbol{y}_{\text{obs}} \in \mathbb{R}^m$. Bayes' Theorem then explicitly computes the posterior pdf as

$$\pi_{\text{post}}(\boldsymbol{x}|\boldsymbol{y}_{\text{obs}}) \propto \pi_{\text{prior}}(\boldsymbol{x})\pi_{\text{like}}(\boldsymbol{y}_{\text{obs}}|\boldsymbol{x}).$$

We choose the prior distribution to be Gaussian, with a covariance operator defined by the square of the inverse of an elliptic PDE operator. This choice yields several benefits. First, it enables implicit representation of the prior covariance operator as (the inverse of) a sparse operator, as opposed to traditional approaches that either store a dense covariance matrix or its approximation by principle vectors. Second, since the covariance operator

is never needed explicitly—only its action on a vector is required— we are able to capitalize on fast $O(n)$ parallel elliptic solvers (in this paper, algebraic multigrid) to form this action via two elliptic solves. Third, the action of the symmetric square root factorization of the prior covariance is available explicitly (via one elliptic solve instead of two). Finally, this choice of covariance is useful for technical reasons, as it guarantees that samples from the prior distribution will be continuous.

The difference between the observables predicted by the model and the actual observations $\boldsymbol{y}_{\text{obs}}$ is due to both measurement and model errors, and is represented by the i.i.d. Gaussian random variable "noise" vector $\boldsymbol{e}$,

$$\boldsymbol{e} = \boldsymbol{y}_{\text{obs}} - \boldsymbol{f}(\boldsymbol{x}),$$

where $\boldsymbol{f}(\boldsymbol{x}) \in \mathbb{R}^m$ is the (generally nonlinear) operator mapping model parameters to output observables. Then the pdf's for the prior and noise can be written in the form

$$\pi_{\text{prior}}(\boldsymbol{x}) \propto \exp\left(-\frac{1}{2}(\boldsymbol{x} - \bar{\boldsymbol{x}}_{\text{prior}})^T \boldsymbol{\Gamma}_{\text{prior}}^{-1}(\boldsymbol{x} - \bar{\boldsymbol{x}}_{\text{prior}})\right),$$

and

$$\pi_{\text{noise}}(\boldsymbol{e}) \propto \exp\left(-\frac{1}{2}(\boldsymbol{e} - \bar{\boldsymbol{e}})^T \boldsymbol{\Gamma}_{\text{noise}}^{-1}(\boldsymbol{e} - \bar{\boldsymbol{e}})\right),$$

respectively, where $\bar{\boldsymbol{x}}_{\text{prior}}$ is the mean of the prior distribution, $\bar{\boldsymbol{e}}$ is the mean of the Gaussian noise, $\boldsymbol{\Gamma}_{\text{prior}} \in \mathbb{R}^{n \times n}$ is the covariance matrix for the prior, and $\boldsymbol{\Gamma}_{\text{noise}} \in \mathbb{R}^{m \times m}$ is the covariance matrix for the noise. Restating Bayes'

theorem with these Gaussian pdf's, we find that the statistical solution of the inverse problem, $\pi_{\text{post}}(\boldsymbol{x})$, is given by

$$\pi_{\text{post}}(\boldsymbol{x}) \propto \exp\Big(-\frac{1}{2}\|\boldsymbol{x} - \bar{\boldsymbol{x}}_{\text{prior}}\|^2_{\boldsymbol{\Gamma}^{-1}_{\text{prior}}} - \frac{1}{2}\|\boldsymbol{y}_{\text{obs}} - \boldsymbol{f}(\boldsymbol{x}) - \bar{\boldsymbol{e}}\|^2_{\boldsymbol{\Gamma}^{-1}_{\text{noise}}}\Big), \tag{3.1}$$

Note that the seemingly simple expression $\boldsymbol{f}(\boldsymbol{x})$ belies the complexity of the underlying computations, which involve: (1) construction of the PDE model for given parameters $\boldsymbol{x}$; (2) solution of the governing PDE model to yield the output state variables; and (3) extraction of the observables from the states at the observation locations in space and time. In §3.7, we provide expressions for the underlying mathematical operators for our target inverse seismic wave propagation problem, in which the parameters are wave speeds in the earth, the governing PDEs describe acoustic wave propagation, and the observations are of velocity waveforms at seismometer locations on earth's surface. In general, $\boldsymbol{f}(\boldsymbol{x})$ is nonlinear, even when the forward PDEs are linear in the state variables (as is the case for the seismic inverse problem), since the model parameters couple with the states nonlinearly in the forward PDEs.

As is clear from the expression (3.1), despite the choice of Gaussian prior and noise probability distributions, the posterior probability distribution need not be Gaussian, due to the nonlinearity of $\boldsymbol{f}(\boldsymbol{x})$. The non-Gaussianity of the posterior poses challenges for computing statistics of interest for typical large-scale inverse problems, since as mentioned in §3.1, π_{post} is often a surface in high dimensions (millions, in our target problem in §3.7), and evaluating

each point on this surface requires the solution of the forward PDEs (wave propagation equations with $O(10^9)$ unknowns, in the target problem). Numerical quadrature to compute the mean and covariance matrix, for example, is completely out of the question. The method of choice for computing statistics is Markov chain Monte Carlo (MCMC), which judiciously samples the posterior distribution, so that sample statistics can be computed. But the use of MCMC for large-scale inverse problems is still prohibitive for expensive forward problems and high dimensional parameter spaces, since even for modest numbers of parameters, the number of samples required can be in the millions. An alternative approach based on linearizing the parameter-to-observable map is discussed next.

3.3 Posterior mean approximation

The mean of the posterior distribution $\bar{\boldsymbol{x}}_{\text{post}}$ can be approximated by finding the point that maximizes the posterior pdf, i.e., the MAP point,

$$\bar{\boldsymbol{x}}_{\text{post}} \approx \boldsymbol{x}_{\text{MAP}} := \arg\max_{\boldsymbol{x}} \pi_{\text{post}}(\boldsymbol{x}).$$

This approximation is exact when the map from parameters to observables, $\boldsymbol{f}(\boldsymbol{x})$, is linear. Finding the MAP point is equivalent to minimizing the negative log of the posterior pdf, i.e.,

$$\bar{\boldsymbol{x}}_{\text{post}} \approx \arg\min_{\boldsymbol{x}} V(\boldsymbol{x}), \tag{3.2}$$

where

$$V(\boldsymbol{x}) = \frac{1}{2}\|\boldsymbol{y}_{\text{obs}} - \boldsymbol{f}(\boldsymbol{x}) - \bar{\boldsymbol{e}}\|^2_{\boldsymbol{\Gamma}^{-1}_{\text{noise}}} + \frac{1}{2}\|\boldsymbol{x} - \bar{\boldsymbol{x}}_{\text{prior}}\|^2_{\boldsymbol{\Gamma}^{-1}_{\text{prior}}}. \tag{3.3}$$

Approximating the mean of the posterior distribution by finding the MAP point is thus equivalent to solving a regularized deterministic inverse problem, where $\boldsymbol{\Gamma}_{\text{prior}}^{-1}$ plays the role of the regularization operator, and $\boldsymbol{\Gamma}_{\text{noise}}^{-1}$ is a weighting for the data misfit term.

Here, we solve the nonlinear least squares optimization problem (3.2) with a parallel inexact Newton–conjugate gradient method. The method requires the computation of gradients and Hessian-vector products of $V(\boldsymbol{x})$ (for which expressions are provided in §3.7 in the context of the seismic inverse problem we target). Rather than provide a detailed description of the method here, we refer to our earlier work presented at SC2002 [4] and SC2003 [3] on parallel scalability of the method, as well as the recent work [67] that includes additional refinements. The main ingredients of the method are:

- inexact Hessian matrix-free Gauss-Newton-conjugate gradient (CG) minimization;
- preconditioning by $\boldsymbol{\Gamma}_{\text{prior}}^{-1}$, carried out by multigrid V-cycles on the underlying elliptic operators;
- Armijo-type backtracking line search globalization;
- computation of gradients of $V(\boldsymbol{x})$ and products of Hessians of $V(\boldsymbol{x})$ with vectors at each CG iteration expressed as solutions of forward and (backward-in-time) adjoint PDEs and their linearizations, all of which inherit the parallel scalability properties of the forward PDE solver;
- algorithmic checkpointing to implement the composition of forward-in-time forward PDE solutions and backward-in-time adjoint PDE solutions

to form gradients without having to store the entire state variable time history; and

- parallel implementation of all components of the method, which are dominated by solution of forward and adjoint-PDEs and evaluation of inner product-like quantities to compose gradient and Hessian-vector quantities.

What can be said about parallel and algorithmic scalability of this method? Because the dominant components of the method can be expressed as solutions or evaluations of PDE-like systems, parallel scalability—that is, maintaining high parallel efficiency as the number of cores increases—is assured whenever a scalable solver for the underlying PDEs is available (which is the case for our target seismic wave propagation problem [201]). The remaining ingredient to obtain overall scalability is that the method exhibit algorithmic scalability, that is with increasing problem size. This is indeed the case: for a wide class of nonlinear inverse problems, the outer Newton iterations and the inner CG iterations are independent of the mesh size (as is the case for our target inverse wave propagation problem, [67]). This is a consequence of the use of a Newton solver, of the compactness of the Hessian of the data misfit term (i.e., the first term on the right hand side of (3.3), as proven for the inverse wave propagation setting in [30]), and the choice of prior preconditioning so that the resulting preconditioned Hessian is a compact perturbation of the identity, for which CG exhibits mesh-independent iterations. Thus, solving the least squares optimization problem (3.2) to approximate the mean of the

posterior distribution by the method outlined above exhibits both parallel and algorithmic—and thus overall—scalability.

As stated above, the focus of this paper is not on the computation of the posterior mean $\bar{\boldsymbol{x}}_{\text{post}}$, but on the significantly more challenging task of characterizing the uncertainty in the mean via computation of the posterior covariance matrix, $\boldsymbol{\Gamma}_{\text{post}} \in \mathbb{R}^{n\times n}$. Linearizing the parameter-to-observable map at the MAP point gives

$$\boldsymbol{f}(\boldsymbol{x}) \approx \boldsymbol{A}(\boldsymbol{x} - \boldsymbol{x}_{\text{MAP}}) + \boldsymbol{f}(\boldsymbol{x}_{\text{MAP}}),$$

where $\boldsymbol{A} \in \mathbb{R}^{m\times n}$ is the Jacobian matrix of $\boldsymbol{f}(\boldsymbol{x})$ evaluated at $\boldsymbol{x}_{\text{MAP}}$. Manipulation of (3.1) shows that $\boldsymbol{\Gamma}_{\text{post}}$ is given by the inverse of the Hessian matrix of the function $V(\boldsymbol{x})$ in (3.3), i.e.,

$$\boldsymbol{\Gamma}_{\text{post}} = \left(\boldsymbol{A}^T\boldsymbol{\Gamma}_{\text{noise}}^{-1}\boldsymbol{A} + \boldsymbol{\Gamma}_{\text{prior}}^{-1}\right)^{-1}. \tag{3.4}$$

In summary, under the assumptions of this section (additive Gaussian noise, Gaussian prior, and linearized parameter-to-observable map), solution of the Bayesian inverse problem is reduced to the characterization of the (Gaussian) posterior distribution $\mathcal{N}(\bar{\boldsymbol{x}}_{\text{MAP}}, \boldsymbol{\Gamma}_{\text{post}})$, where $\boldsymbol{\Gamma}_{\text{post}}$ is the inverse of the Hessian of $V(\boldsymbol{x})$ at $\boldsymbol{x}_{\text{MAP}}$.

The primary difficulty here is that the large parameter dimension n prevents any representation of the posterior covariance $\boldsymbol{\Gamma}_{\text{post}}$ as a dense operator. In particular, the Jacobian of the parameter-to-observable map, $\boldsymbol{A}$, is formally a dense matrix, and requires n forward PDE solves to construct.

This is intractable when n is large and the PDEs are expensive, as in our case. However, a key feature of the operator $\boldsymbol{A}$ is that its action on a (parameter field-like) vector can be formed by solving a (linearized) forward PDE problem; similarly, the action of its transpose $\boldsymbol{A}^T$ on a (observation-like) vector can be formed by solving a (linearized) adjoint PDE. Explicit expressions for these operations will be given for our specific target inverse problem in §3.7. In the next two sections, we present algorithms that exploit this property, as well as the spectral decay of the data misfit Hessian, to approximate the posterior covariance matrix with controlled accuracy at a cost that is independent of the parameter dimension.

3.4 Posterior covariance approximation

For many ill-posed inverse problems, the Hessian matrix of the data misfit term in (3.3), defined as

$$\boldsymbol{H}_{\text{misfit}} \overset{\text{def}}{=} \boldsymbol{A}^T \boldsymbol{\Gamma}_{\text{noise}}^{-1} \boldsymbol{A}, \tag{3.5}$$

is a discretization of a compact operator, i.e., its eigenvalues collapse to zero. This can be understood intuitively, since only the modes of the parameter field that strongly influence the observations (through the linearized parameter-to-observable map $\boldsymbol{A}$) will be present in the dominant spectrum of (3.5). In many ill-posed inverse problems, observations are sparse compared to the parameter dimensions, and numerous modes of the parameter field (for example, highly oscillatory ones) will have negligible effect on the observables. The range space

thus is effectively finite-dimensional even before discretization (and therefore independent of any mesh), and the eigenvalues decay, often rapidly, to zero. In this section, we exploit this low-rank structure to construct scalable algorithms to approximate the posterior covariance operator.

Rearranging the expression for $\boldsymbol{\Gamma}_{\text{post}}$ in (3.4) to factor out $\boldsymbol{\Gamma}_{\text{prior}}^{1/2}$ gives

$$\boldsymbol{\Gamma}_{\text{post}} = \boldsymbol{\Gamma}_{\text{prior}}^{1/2} \left(\boldsymbol{\Gamma}_{\text{prior}}^{1/2} \boldsymbol{A}^T \boldsymbol{\Gamma}_{\text{noise}}^{-1} \boldsymbol{A} \boldsymbol{\Gamma}_{\text{prior}}^{1/2} + \boldsymbol{I} \right)^{-1} \boldsymbol{\Gamma}_{\text{prior}}^{1/2}. \tag{3.6}$$

This factorization exposes the *prior-preconditioned Hessian of the data misfit*,

$$\tilde{\boldsymbol{H}}_{\text{misfit}} \stackrel{\text{def}}{=} \boldsymbol{\Gamma}_{\text{prior}}^{1/2} \boldsymbol{A}^T \boldsymbol{\Gamma}_{\text{noise}}^{-1} \boldsymbol{A} \boldsymbol{\Gamma}_{\text{prior}}^{1/2}. \tag{3.7}$$

In the next section we present a randomized algorithm to construct a low rank approximation of this matrix at a cost (in PDE solves) that is independent of the parameter dimension (compared to n PDE solves to construct the full matrix). In this section, we assume only that such a low rank construction is possible. Let λ_i and $\boldsymbol{v}_i$ be the eigenvalues and eigenvectors of $\tilde{\boldsymbol{H}}_{\text{misfit}}$. Let $\boldsymbol{\Lambda} = \text{diag}(\lambda_i) \in \mathbb{R}^{n \times n}$ be the diagonal matrix of its eigenvalues, and define as $\boldsymbol{V} \in \mathbb{R}^{n \times n}$ the matrix whose columns are the eigenvectors $\boldsymbol{v}_i$ of $\tilde{\boldsymbol{H}}_{\text{misfit}}$. Then replace $\tilde{\boldsymbol{H}}_{\text{misfit}}$ by its spectral decomposition:

$$\left(\boldsymbol{\Gamma}_{\text{prior}}^{1/2} \boldsymbol{A}^T \boldsymbol{\Gamma}_{\text{noise}}^{-1} \boldsymbol{A} \boldsymbol{\Gamma}_{\text{prior}}^{1/2} + \boldsymbol{I} \right)^{-1} = (\boldsymbol{V} \boldsymbol{\Lambda} \boldsymbol{V}^T + \boldsymbol{I})^{-1}. \tag{3.8}$$

When the eigenvalues of $\tilde{\boldsymbol{H}}_{\text{misfit}}$ decay rapidly, we can extract a low-rank approximation of $\tilde{\boldsymbol{H}}_{\text{misfit}}$ by retaining only the r largest eigenvalues and corresponding eigenvectors,

$$\boldsymbol{\Gamma}_{\text{prior}}^{1/2} \boldsymbol{A}^T \boldsymbol{\Gamma}_{\text{noise}}^{-1} \boldsymbol{A} \boldsymbol{\Gamma}_{\text{prior}}^{1/2} \approx \boldsymbol{V}_r \boldsymbol{\Lambda}_r \boldsymbol{V}_r^T.$$

Here $\boldsymbol{V}_r \in \mathbb{R}^{n\times r}$ contains only the r eigenvectors of $\tilde{\boldsymbol{H}}_{\text{misfit}}$ that correspond to the r largest eigenvalues, which are assembled in the diagonal matrix $\boldsymbol{\Lambda}_r = \text{diag}(\lambda_i) \in \mathbb{R}^{r\times r}$. To obtain the posterior covariance matrix, we employ the Sherman-Morrison-Woodbury formula to perform the inverse in (3.6),

$$\left(\boldsymbol{\Gamma}_{\text{prior}}^{1/2}\boldsymbol{A}^T\boldsymbol{\Gamma}_{\text{noise}}^{-1}\boldsymbol{A}\boldsymbol{\Gamma}_{\text{prior}}^{1/2} + \boldsymbol{I}\right)^{-1} = \boldsymbol{I} - \boldsymbol{V}_r\boldsymbol{D}_r\boldsymbol{V}_r^T + \mathcal{O}\left(\sum_{i=r+1}^{n}\frac{\lambda_i}{\lambda_i+1}\right),$$

where $\boldsymbol{D}_r \overset{\text{def}}{=} \text{diag}(\lambda_i/(\lambda_i+1)) \in \mathbb{R}^{r\times r}$. The last term in the expression above shows the error due to truncation in terms of the discarded eigenvalues; this provides a criterion for truncating the spectrum, namely r is chosen such that λ_r is small relative to 1. With this low-rank approximation, the final expression for the approximate posterior covariance follows from (3.6),

$$\boldsymbol{\Gamma}_{\text{post}} \approx \boldsymbol{\Gamma}_{\text{prior}} - \boldsymbol{\Gamma}_{\text{prior}}^{1/2}\boldsymbol{V}_r\boldsymbol{D}_r\boldsymbol{V}_r^T\boldsymbol{\Gamma}_{\text{prior}}^{1/2}. \tag{3.9}$$

Note that (3.9) expresses the posterior uncertainty (in the form of a covariance matrix) as the prior uncertainty, less any information gained from the data, filtered through the prior.

3.5 A randomized algorithm for low-rank Hessian approximation

We now address the construction of the low rank approximation of $\tilde{\boldsymbol{H}}_{\text{misfit}}$ that was invoked in the previous section. As argued above, the data inform only a limited number of modes of the parameter field, resulting in a

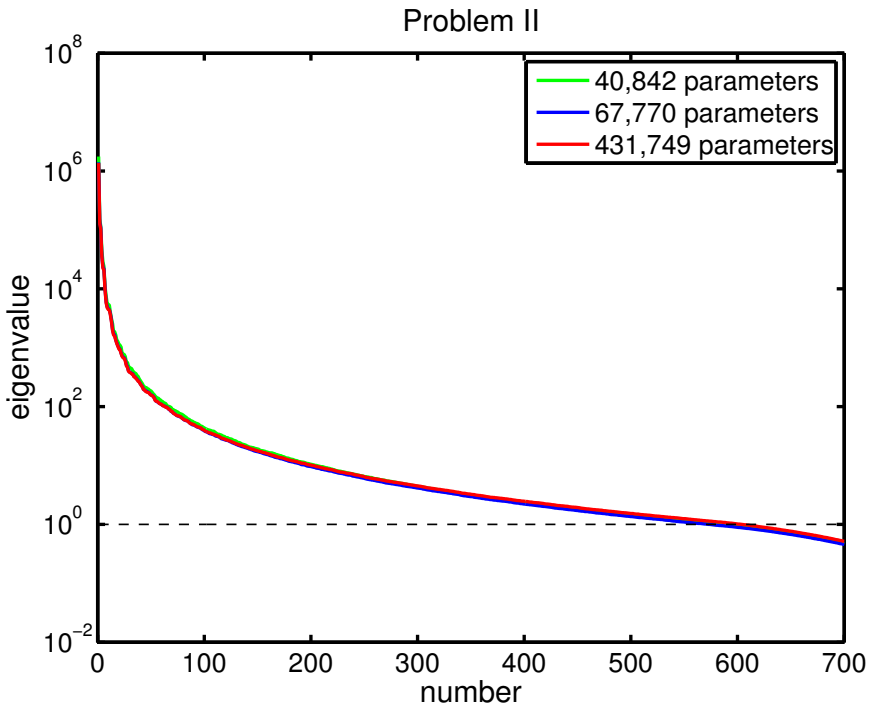

Figure 3.1: Log-linear plot of the spectrum of prior-preconditioned data misfit Hessian ($\tilde{\boldsymbol{H}}_{\text{misfit}}$) for three successively finer parameter meshes of an inverse wave propagation problem [33]. The spectra lie on top of each other, indicating mesh independence (and therefore parameter-dimension independence) of the low rank approximation. The eigenvalues are truncated when they are small relative to 1, which in this case results in retaining between 0.2 and 2% of the spectrum.

data misfit Hessian matrix that admits a low rank representation. This is observed numerically (see Figure 3.1) and has recently been proven theoretically in several settings [29, 30]. Moreover, preconditioning with the prior operator as in (3.7) further filters out modes of the parameter space that are already well-determined from prior knowledge (i.e., a smoothing prior such as the one we employ here assigns low probability to highly oscillatory modes.)

We exploit this structure to construct a low rank approximation of

$\tilde{\boldsymbol{H}}_{\text{misfit}}$ using randomized algorithms for approximate matrix decomposition [98, 134]. Their performance is comparable to Krylov methods (such as Lanczos) we employed previously [73, 143]. However, they have a significant edge over these deterministic methods for large-scale problems, since the required Hessian matrix-vector products are independent of each other, providing asynchronousity and fault tolerance. Before discussing these advantages, let us summarize the algorithm.

To approximate the spectral decomposition of $\tilde{\boldsymbol{H}}_{\text{misfit}} \in \mathbb{R}^{n\times n}$, we generate a random matrix $\boldsymbol{R} \in \mathbb{R}^{n\times r}$ (r is of the order of the numerical rank of $\tilde{\boldsymbol{H}}_{\text{misfit}}$, so in our case $r \ll n$) with i.i.d. Gaussian entries, and compute the product $\boldsymbol{Y} = \tilde{\boldsymbol{H}}_{\text{misfit}}\boldsymbol{R}$. Since each column vector in $\boldsymbol{R}$ is an independent random vector, the computation of $\boldsymbol{Y}$ decouples into r separate matrix-vector product with $\tilde{\boldsymbol{H}}_{\text{misfit}}$. As can be seen from (3.7), each matrix-vector product requires a pair of forward/adjoint PDE solves (to form actions of $\boldsymbol{A}$ and $\boldsymbol{A}^T$ on vectors), as well as a pair of elliptic operator solves (to form actions of $\boldsymbol{\Gamma}_{\text{prior}}^{1/2}$ on vectors). The latter are much cheaper than the former, in the typical case when the PDE model governing the inverse problem is large scale.

Let $\boldsymbol{Q}$ be an orthonormal basis for $\boldsymbol{Y}$, which approximates the range space of $\tilde{\boldsymbol{H}}_{\text{misfit}}$. Following the "single-pass" approach of [98], we compute the approximation to $\tilde{\boldsymbol{H}}_{\text{misfit}}$ in the basis $\boldsymbol{Q}$:

$$\boldsymbol{B} \stackrel{\text{def}}{=} (\boldsymbol{Q}^T\boldsymbol{Y})(\boldsymbol{Q}^T\boldsymbol{R})^{-1} \approx \boldsymbol{Q}^T\tilde{\boldsymbol{H}}_{\text{misfit}}\boldsymbol{Q}. \tag{3.10}$$

Here $\boldsymbol{B}$, $\boldsymbol{Q}^T\boldsymbol{Y}$, and $(\boldsymbol{Q}^T\boldsymbol{R})^{-1}$ are all matrices of dimension r, which is much

smaller than n, and thus we are able to decompose the symmetric matrix $\boldsymbol{B}$ as $\boldsymbol{Z}\boldsymbol{\Lambda}\boldsymbol{Z}^T$ using dense linear algebra. The dominant vectors of $\tilde{\boldsymbol{H}}_{\text{misfit}}$ are then returned as $\boldsymbol{V} = \boldsymbol{Q}\boldsymbol{Z}$, with eigenvalues on the diagonal of $\boldsymbol{\Lambda}$. Thus, we find the desired decomposition

$$\tilde{\boldsymbol{H}}_{\text{misfit}} \approx \boldsymbol{V}\boldsymbol{\Lambda}\boldsymbol{V}^T. \tag{3.11}$$

Finally, randomized methods also provide an estimate of the spectral norm of $\boldsymbol{I} - \boldsymbol{Q}\boldsymbol{Q}^T\tilde{\boldsymbol{H}}_{\text{misfit}}$, which bounds the error that we make in our low rank approximation. To be precise, the bound derived in [98] is

$$\left\|(\boldsymbol{I} - \boldsymbol{Q}\boldsymbol{Q}^T)\tilde{\boldsymbol{H}}_{\text{misfit}}\right\| \leq \alpha\sqrt{\frac{2}{\pi}}\max_{i=1,\ldots,r}\left\|(\boldsymbol{I} - \boldsymbol{Q}\boldsymbol{Q}^T)\boldsymbol{A}\boldsymbol{\omega}^{(i)}\right\|, \tag{3.12}$$

attained with probability of at least $1 - \alpha^{-r}$, where $\boldsymbol{\omega}^{(i)}$ are vectors with i.i.d. standard normal entries.

To summarize, the construction of a low-rank approximation of $\tilde{\boldsymbol{H}}_{\text{misfit}}$ is dominated by its application to random vectors, which entails a pair of forward/adjoint PDE solves. The independence of these matrix-vector products from each other is of particular importance for problems in which the parameter-to-observable map $\boldsymbol{f}(\boldsymbol{x})$ has to be computed on large parallel supercomputers for the following reasons:

- Cache and memory efficiency: For parameter-to-observable maps that involve the solution of a PDE, the application of the Hessian to multiple vectors requires the solution of (linearized) forward/adjoint PDEs for multiple right-hand sides. Amortizing data movement over the multiple

right-hand sides results in significantly greater memory and cache efficiency than can be obtained with sequential right-hand sides, as required by classical Krylov methods.

- Fault-tolerance: the construction of the low-rank matrix approximation is done as a post-processing step when a sufficient number of matrix-vector products is available. The asynchronous nature of the matrix-vector products provides greater fault tolerance (for example, the low rank approximation in §3.7 was computed using 10 different jobs with different run times and core counts ranging from 32K to 108K.

3.6 Scalability of the UQ method

We now discuss the overall scalability of our UQ method to high-dimensional parameter spaces. First, we summarize the scalability of the construction of the low-rank-based approximate posterior covariance matrix in (3.9). As stated before, the linearized parameter-to-observable map $\boldsymbol{A}$ cannot be constructed explicitly, since it requires n linearized forward PDE solves. However, its action on a vector can be computed by solving a single linearized forward PDE, regardless of the number of parameters n and observations m. Similarly, the action of $\boldsymbol{A}^T$ on a vector can be computed by solving a linearized adjoint PDE. Moreover, the prior is usually much cheaper to apply than the forward or adjoint PDE solution (in our context, it is a single elliptic solve). Therefore, the cost of applying $\tilde{\boldsymbol{H}}_{\text{misfit}}$ to a vector—and thus the per iteration cost of the randomized algorithm of §3.5—is dominated by the solution

of a pair of linearized forward and adjoint PDEs (explicit expressions for this matrix-vector product will be given for the target problem of inverse wave propagation in §3.7).

The remaining component to establish scalability of the low-rank approximation of $\tilde{\boldsymbol{H}}_{\text{misfit}}$ is independence of the rank r—and therefore the number of matrix-vector products, and hence PDE solves—from the parameter dimension n. This is the case when $\boldsymbol{H}_{\text{misfit}}$ in (3.5) is a (discretization of a) compact operator, and when preconditioning by $\boldsymbol{\Gamma}_{\text{prior}}$ does not destroy the spectral decay. This situation is typical for many ill-posed inverse problems, in which the prior is either neutral or of smoothing type (here, we employ a prior that is the inverse of an elliptic operator). Compactness of the data misfit Hessian $\boldsymbol{H}_{\text{misfit}}$ for inverse wave propagation problems has long been observed (e.g., [51]). Recently, we have proven compactness for the inverse wave propagation problem for both continuous and pointwise observation operators for both shape and medium scattering [29, 30]. Specifically, we have shown that the data misfit Hessian is a compact operator. We also quantify the decay of data misfit Hessian eigenvalues in terms of the smoothness of the medium, i.e., the smoother it is, the faster the decay rate. Under some conditions, the rate can be shown to be exponential. That is, the data misfit Hessian can be approximated well with a handful of its dominant eigenvectors and eigenvalues. In conclusion, a low-rank approximation of $\tilde{\boldsymbol{H}}_{\text{misfit}}$ can be made that does not depend on the parameter dimension, and depends only on the information content of the data, filtered through the prior.

Once the r eigenpairs defining the low rank approximation have been computed, estimates of uncertainty can be computed by interrogating $\mathbf{\Gamma}_{\text{post}}$ in (3.9) at a cost of just r inner products (which are negligible) plus elliptic solves representing the action of the square root of the prior $\mathbf{\Gamma}_{\text{prior}}^{1/2}$ on a vector (here carried out with algebraic multigrid and therefore scalable). For example, samples can be drawn from the Gaussian defined with a covariance $\mathbf{\Gamma}_{\text{post}}$, a row/column of $\mathbf{\Gamma}_{\text{post}}$ can be computed, and the action of $\mathbf{\Gamma}_{\text{post}}$ in a given direction can be formed, all at cost that is $O(rn)$ for the inner products in addition to the $O(n)$ cost of the multigrid solve. Moreover, the posterior variance field, i.e., the diagonal of $\mathbf{\Gamma}_{\text{post}}$, can be found with $O(rn)$ linear algebra plus $O(r)$ multigrid solves.

In summary, we have a method for estimating the posterior covariance—and thus the uncertainty in the solution of the linearized inverse problem—that requires a constant number of PDE solves, dependent only on the information content of the data filtered through the prior (i.e., r), but independent of the number of parameters (n), the number of observations (m), and the number of state variables. Moreover, since the dominant cost of the posterior covariance construction is that of solving forward and adjoint-like PDEs, parallel scalability of the overall uncertainty quantification method follows when the forward PDE solver is scalable (this will be demonstrated for the case of our seismic wave propagation solver in the next section).

3.7 Application to global seismic inversion

In recent years, the methodology for scalable parallel solution of forward seismic wave propagation problems on supercomputers by spectral element [43, 121], finite difference [58], finite element [3], and discontinuous Galerkin [36] methods has matured. This motivates our present interest in the seismic inverse problem of determining an earth model from surface observations of seismic waveforms; indeed, we are interested not just in the solution of this inverse problem, but in quantifying the uncertainties in its solution using the method proposed in this paper. In previous sections, our method and underlying algorithms were presented for generic prior and likelihood functions. §3.7.1 provides explicit expressions for these functions (in infinite dimensions) for the specific seismic inverse problem we address, along with explicit expressions for gradient and Hessian-vector products, which are needed for computing the mean and covariance estimates. The latter expressions involve solutions of forward and adjoint wave propagation PDEs and their linearizations. §3.7.2 gives an overview of the forward wave equation solver and provides near-full system strong scalability results on the Jaguar supercomputer at ORNL. §3.7.3 describes the setup of the seismic inverse problem: the configuration of sources and receivers, the generation of synthetic seismogram observations, the choice of prior and noise covariances, the parametrization of wave speed, and the mesh generation. §3.7.3 presents results on quantifying uncertainties in the solution of a linearized global seismic inverse problem characterized by one million uncertain parameters. This is the largest—in fact the first—solution

of which we are aware of a statistical inverse problem whose forward solver has required a supercomputer, made possible because of the parameter-dimension-independent scaling of our method.

3.7.1 Posterior and its derivatives

In this section we give explicit expressions for $V(\boldsymbol{x})$, the negative log of the posterior pdf for the seismic inverse problem we target, along with expressions for its gradient and Hessian-vector product. The expressions are written in strong, infinite-dimensional form, for clarity. The inversion parameter is taken as $c = c(\boldsymbol{x})$, the local acoustic wave speed of the medium. We can write the negative log posterior as

$$\mathcal{V}(c) := \frac{1}{2} \left\| \mathcal{B}\boldsymbol{v}(c) - \boldsymbol{v}^{\text{obs}} \right\|^2_{\Gamma^{-1}_{\text{noise}}} + \frac{1}{2} \left\| c - \bar{c} \right\|^2_{\Gamma^{-1}_{\text{prior}}},$$

where the data misfit (the first term) is a finite dimensional norm due to the pointwise observations in time and space, and the prior term (the second term) is an infinite dimensional norm, with the elliptic prior operator $\Gamma^{-1}_{\text{prior}}$ taken as an anisotropic biharmonic. The wave propagation variables—the velocity vector $\boldsymbol{v}$ and the trace of the strain tensor e (i.e., the dilation) depend on c through the solution of the forward wave propagation equations (written in

first-order form):

$$\begin{aligned}
\rho \boldsymbol{v}_t - \nabla(\rho c^2 e) &= \boldsymbol{g} \ \text{ in } \ \Omega \times (0,T), \\
e_t - \nabla \cdot \boldsymbol{v} &= 0 \ \text{ in } \ \Omega \times (0,T), \\
\rho \boldsymbol{v} = \boldsymbol{0}, e &= 0 \ \text{ in } \ \Omega \times \{t = 0\}, \\
e &= 0 \ \text{ on } \ \partial\Omega \times (0,T).
\end{aligned}$$

Here, ρ and $\boldsymbol{g}$ are known density and seismic source, $\boldsymbol{v}^{\text{obs}}$ are observations at receivers, $\mathcal{B}$ is an observation operator, and Γ_{prior} and Γ_{noise} are the prior and noise covariance operators.

The adjoint approach allows us to write the gradient at a given point c in parameter space as

$$\mathcal{G}(c) := 2\rho c \int_0^T e(\nabla \cdot \boldsymbol{w})\, dt + \Gamma_{\text{prior}}^{-1}(c - \bar{c}),$$

where the adjoint velocity $\boldsymbol{w}$ and adjoint strain dilation d satisfy the *adjoint wave propagation equations*

$$\begin{aligned}
-\rho \boldsymbol{w}_t + \nabla(c^2 \rho d) &= -\mathcal{B}^* \Gamma_{\text{noise}}^{-1}(\mathcal{B}\boldsymbol{v} - \boldsymbol{v}^{\text{obs}}) \text{ in } \Omega \times (0,T), \\
-d_t + \nabla \cdot \boldsymbol{w} &= 0 \qquad\qquad \text{in } \Omega \times (0,T), \\
\rho \boldsymbol{w} = \boldsymbol{0}, d &= 0 \qquad\qquad \text{in } \Omega \times \{t = T\}, \\
d &= 0 \qquad\qquad \text{on } \Gamma \times (0,T).
\end{aligned}$$

The adjoint wave equations are reversed in time and have the data misfit as source term, but otherwise resemble the forward wave equations.

Similarly, the action of the Hessian operator in the direction $\tilde{c}$ at a point c is given by

$$\mathcal{H}(c)\tilde{c} := 2\rho \int_0^T ce(\nabla \cdot \tilde{\boldsymbol{w}}) + c\tilde{e}(\nabla \cdot \boldsymbol{w}) + \tilde{c}e(\nabla \cdot \boldsymbol{w})\, dt + \Gamma_{\text{prior}}^{-1}\tilde{c},$$

where $\tilde{\boldsymbol{v}}$ and $\tilde{e}$ satisfy the *incremental forward wave propagation equations*

$$\begin{aligned}
\rho \boldsymbol{v}_t - \nabla(\rho c^2 \tilde{e}) &= \nabla(2\rho c \tilde{c} e) && \text{in } \Omega \times (0,T),\\
e_t - \nabla \cdot \tilde{\boldsymbol{v}} &= 0 && \text{in } \Omega \times (0,T),\\
\rho \tilde{\boldsymbol{v}} = \boldsymbol{0}, \tilde{e} &= 0 && \text{in } \Omega \times \{t = 0\},\\
\tilde{e} &= 0 && \text{on } \Gamma \times (0,T).
\end{aligned}$$

On the other hand, $\tilde{\boldsymbol{w}}$ and $\tilde{d}$ satisfy the *incremental adjoint wave propagation equations*

$$\begin{aligned}
-\rho \boldsymbol{w}_t + \nabla(c^2 \rho \tilde{d}) &= -\nabla(2\tilde{c} c \rho d) - \mathcal{B}^* \Gamma_{\text{noise}}^{-1} \mathcal{B} \tilde{\boldsymbol{v}} && \text{in } \Omega \times (0,T),\\
-d_t + \nabla \cdot \tilde{\boldsymbol{w}} &= 0 && \text{in } \Omega \times (0,T),\\
\rho \tilde{\boldsymbol{w}} = \boldsymbol{0}, \tilde{d} &= 0 && \text{in } \Omega \times \{t = T\},\\
\tilde{d} &= 0 && \text{on } \Gamma \times (0,T).
\end{aligned}$$

The incremental forward and incremental adjoint wave equations are seen to be linearized versions of their forward and adjoint counterparts, and thus differ only in the source terms.[2]

[2]The infinite dimensional expressions for the gradient and Hessian action given above are actually not consistent with the discrete gradient and Hessian-vector product obtained by first discretizing the negative log posterior and wave equation and then differentiating with respect to parameters. Additional jump terms at element interfaces due to the dG discretization appear; in our implementation, we include these terms to insure consistency with discrete counterparts.

Thus, we see that computation of gradients (as needed in the posterior mean approximation) and Hessian actions on vectors (as needed in the posterior covariance approximation) amount to solution of a pair of forward/adjoint wave equations each.

3.7.2 Wave propagation solver and its strong scalability

The forward wave equation, and its three variants (adjoint, incremental forward, incremental adjoint) described in the previous section, are solved using a high-order discontinuous Galerkin (dG) method. Details on the forward solver are provided in [201]; here we summarize the salient features:

- discretization that supports h-nonconforming hexahedral elements on a 2:1 balanced forest-of-octrees mesh;
- an element basis that is a tensor product of Lagrange polynomials of arbitrarily high degree based on the Legendre-Gauss-Lobatto (LGL) nodes;
- LGL numerical quadrature, which produces a diagonal mass matrix;
- solution of the Riemann problem at material interfaces (elastic-elastic, elastic-acoustic, acoustic-acoustic);
- mortar-based implementation of flux on 2:1 nonconforming faces;
- time integration by classical four-stage fourth-order Runge Kutta;
- guaranteed consistency, semi-discrete stability, and optimal order convergence for non-conforming meshes [28].

To model global seismic wave propagation, we model the earth as a sphere with a radius of 6,371 km, where the speed of acoustic (pressure) waves

varies throughout the domain. To generate the finite element mesh, we decompose the earth into 13 warped cubes. The inner core comprises one central cube, surrounded by two layers of six additional cubes. Each cube is the root of an adaptive octree, which can be arbitrarily refined, thus creating a mesh of curved hexahedral elements. The mesh is aligned to the interface between the outer core and the mantle, and several weaker discontinuities between layers, and refined locally to resolve varying seismic wavelengths up to a target frequency. The wave speed $c(\boldsymbol{x})$ is approximated with piecewise trilinear finite elements, and the wave equation variables (velocity and strain) are discretized using high-order (spectral) discontinuous Galerkin finite elements on the same hexahedral mesh. For the distributed storage and adaptation of both the parameter and wave propagation meshes, we use our `p4est` library of fast forest-of-octree algorithms for scalable adaptive mesh refinement, which have been shown to scale to over 220,000 CPU cores and impose minimal computational overhead [36, 37]. The time spent in meshing is insignificant relative to that of numerical solution of the wave equation.

The central difficulty of UQ is its need for repeated solution of the governing PDE model, in our case the wave propagation equations. Conventional sampling methods will take millions of wave propagation solutions (realistically, much more) to explore the posterior distribution for the million-parameter problem we solve in this section. For the frequencies we target, a single wave propagation solve takes a minute on 64K Jaguar cores; conventional sampling methods are thus out of the question. The low-rank Hessian-based

method we have presented here, which captures and exploits the local structure of the posterior in the directions informed by the data by computing curvature information based on additional wave equations (adjoint and incremental forward and adjoint), reduces the number of wave propagation solutions by orders of magnitude. Still, thousands of wave equation solves are needed, and we must use all available computing resources. As a result, we insist on excellent strong scalability of our wave equation solver to achieve acceptable time-to-solution. Taken together, the high-order discretization, discontinuous elements, explicit RK scheme, and space filling curve partitioning underlying our forest-of-octree mesh data structure should yield excellent scalability; indeed, we have shown near ideal parallel efficiency in weak scaling on up to 220,000 cores of the Jaguar system at ORNL [201]. Here, we investigate the extreme limits of strong scaling to determine how fine a granularity one can employ in the repeated wave solutions. Table 3.1 shows that our wave equation solver exhibits excellent strong scaling over a wide range of core counts. These results are significant, since we are using just third-order elements (higher order creates more work per element, relative to data movement). For the large problem, for example we maintain 71% parallel efficiency in strong scaling from 1024 to 262,144 cores. The largest core count problem has just 62 elements per core.

Table 3.1: Strong scaling of the forward solver

#cores	time [ms]	elem/core	efficiency [%]
256	1630.80	4712	100.0
512	832.46	2356	98.0
1024	411.54	1178	99.1
8192	61.69	148	82.6
65536	11.79	19	54.0
131072	7.09	10	44.9
262144	4.07	5	39.2
1024	5423.86	15817	100.0
4096	1407.81	3955	96.3
8192	712.91	1978	95.1
16384	350.43	989	96.7
32768	211.86	495	80.0
65536	115.37	248	73.5
131072	57.27	124	74.0
262144	29.69	62	71.4

Strong scaling results on ORNL's Jaguar XK6 system for global seismic wave propagation solutions for two problem sizes. We report the time per time step in milliseconds on meshes with 1,206,050 (upper table) and 16,195,864 (lower table) 3rd order discontinuous Galerkin finite elements, corresponding to 694 million and 9.3 billion spatial degrees of freedom, respectively. The elem/core column reports the maximum number of elements owned by any core. For strong scaling from 256 to 262,144 cores, the parallel efficiency is still as high as 39% for the small problem. For the larger problem and a 256-fold increase in problem size, we find a parallel efficiency of 71%. At 262,144 cores, each core owns just 4 or 5 elements for the small problem, and 61 or 62 elements for the larger problem. The larger run sustains a double precision floating point rate of 111 teraflops per second (based on performance counters from the *PAPI* library [1].

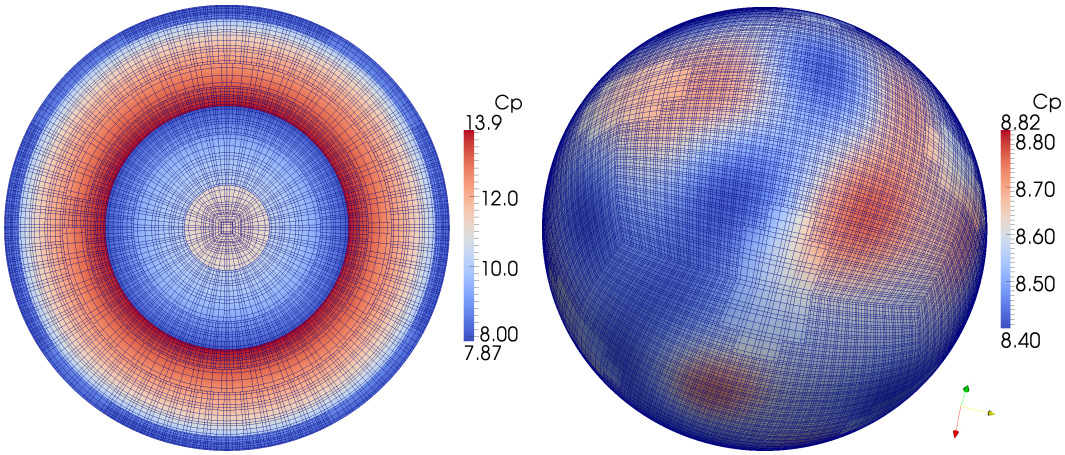

Figure 3.2: (Coarser version of) mesh used for the wave propagation simulation and "true" pressure wave speed c in km/s. Left: section through earth model. Right: surface at depth of 222 km showing lateral variations of up to 7%. Wave propagation mesh is tailored to the local seismic wave lengths.

3.7.3 Inverse problem solution and its uncertainty

This section presents solution of the statistical inverse problem. First we define the inverse problem setup. Both the prior mean and the initial guess for the iterative solution of the nonlinear least squares optimization problem (3.2) (to find the MAP point) are derived from the radially symmetric preliminary reference earth model (PREM) [63], which dates to 1981. We take the "true" earth to be given by the more recent S20RTS velocity model (converted from shear to acoustic wave speed anomaly) [196], which superposes lateral wave speed variations on PREM, as seen in Figure 3.2. Synthetic observations are generated from solution of the wave equation for an S20RTS earth model, with seismic sources at the North pole and at 90° intervals along

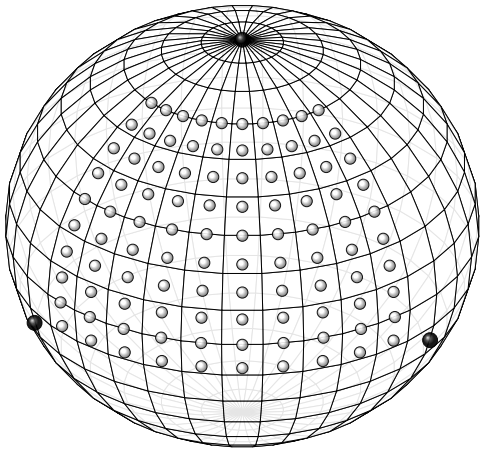

Figure 3.3: Location of five simultaneous seismic sources (black spheres; two in back not visible) and 100 receivers (white spheres).

the equator, all of them at a depth of 10 km. All five point sources are taken to occur simultaneously. A total of 100 receivers in the Northern and Eastern hemispheres are distributed along zonal lines at 10° spacing. The source and receiver configuration is illustrated in Figure 3.3. The observations consist of the first 61 Fourier coefficients of the Fourier-transformed seismogram (time history of ground motion) at each receiver location. The noise distribution for these data is taken as i.i.d. Gaussian with mean zero and a standard deviation of 9.34×10^{-3}.

We use a 3rd-order discontinuous finite elements mesh to resolve seismic wavelengths corresponding to a source with maximum frequency of 0.07 Hz. This requires a mesh with 1,093,784 elements, which leads to 630 million wave propagation spatial unknowns (velocity and strain) for the forward problem, and 1,067,050 unknown wave speed parameters for the statistical inverse prob-

lem. The observation time window for the inverse problem is 1,000 seconds, which leads to 2400 discrete time steps. This simulation time is sufficient for the waves to travel about two-thirds of the earth's diameter. A single wave solve takes about one minute on 64K Jaguar cores. As discussed in §3.7.1, two wave solves are needed in each gradient or Hessian-vector computation. However, since these expressions combine wave equation solutions in opposite time direction, the work-optimal choice of solving two wave equations requires storage of the entire time history, which is prohibitive. Instead, we use algorithmic checkpointing methods, which cut the necessary storage but increases the number of wave propagation solutions to five per Hessian-vector product (two forward, two incremental forward, and one adjoint solve) [67]. Thus, a single Hessian-vector product takes about 5 minutes on 65K Jaguar cores.

The posterior mean is approximated by solving the nonlinear least squares optimization problem (3.2) to find the MAP point, using the inexact Gauss Newton-CG method described in §3.2, initialized with the prior mean (the PREM model), and terminated after 3 orders of magnitude reduction in the gradient, which was achieved after a total of 320 CG iterations (summed across Newton iterations). A comparison of the approximate mean with the "true" earth model (S20RTS) is displayed in Figure 3.4. The MAP solution is seen to resemble the "true" parameter field well in the Northern hemisphere, which has good receiver coverage.

We approximate the covariance matrix at the MAP point via a low-rank representation employing 488 products of the Hessian matrix with random vec-

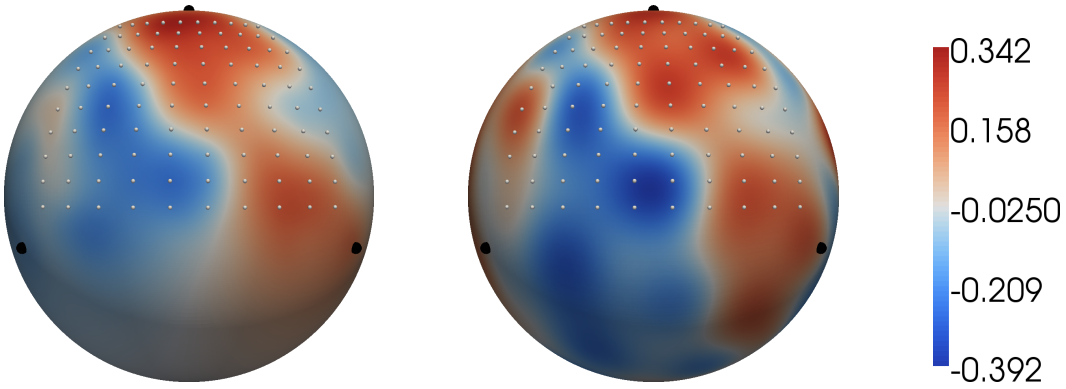

Figure 3.4: Comparison of MAP of posterior pdf (left) with the "true" earth model (right) at a depth of 67 km. Source locations are indicated with black spheres and seismic receiver stations are indicated by white spheres.

tors. The effective problem dimension is thus reduced from 1.07 million to 488, a factor of over 2000 reduction. Figure 3.5 depicts the first 488 eigenvalues of the million-dimensional parameter field, indicating the rapid decay in information content of the data, a fact that we exploit to make the UQ problem tractable.

The reduction in the variance between prior and posterior due to the information (about the earth model) content of the data—i.e., the diagonal of the second term in (3.9), the expression for the posterior covariance—is shown in Figure 3.7. We observe that in the region where sensors are placed (the visible portion of the Northern hemisphere), we get a large reduction in variance due to the data. In regions where there are no sensors, the reduction in variance is substantially less. Additionally, Figure 3.8 displays the variance reduction on a slice through the equator of the earth, and we again see that

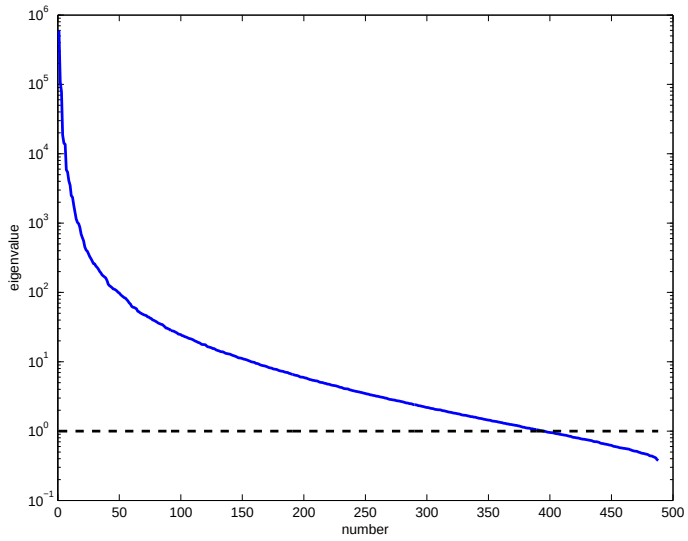

Figure 3.5: Logarithmic plot of the spectrum of prior-preconditioned data misfit Hessian.

the largest variance reduction (depicted in red) is achieved near the surface where the sensors are located, although some reduction is also achieved well into the earth's mantle. Finally, Figure 3.6 shows samples from the prior and the posterior pdf; the difference between the two sets of samples reflects the information gained from the data in solving the inverse problem. Note the regions of large variability in the posterior samples, which reflect the absence of receivers.

3.8 Conclusions

We have addressed UQ for large-scale inverse problems. We adopt the Bayesian inference framework: given observational data and their uncertainty, the governing forward problem and its uncertainty, and a prior probability

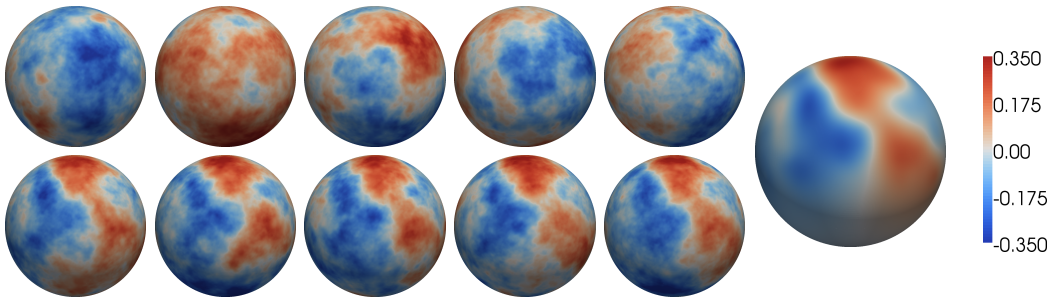

Figure 3.6: Samples from the prior (top row) and posterior (bottom row) distributions. The difference between the prior and posterior samples reflects the information (about the earth model) learned from the data. The large scale features of the posterior samples consistently resemble the posterior mean (right). The fine scale features however are not expected to be influenced by the data, and qualitatively resemble the fine scale features of the prior samples. Note the small variability across samples in the Northern hemisphere—reflecting the receiver coverage there—while the Southern hemisphere exhibits large variability in the inferred model, reflecting that uncertainty due to the lack of receivers.

distribution describing uncertainty in the parameters, find the posterior probability distribution over the parameters. The posterior pdf is a surface in high dimensions, and the standard approach is to sample it via a Markov-chain Monte Carlo (MCMC) method and then compute statistics of the samples. However, the use of conventional MCMC methods becomes intractable for high dimensional parameter spaces and expensive-to-solve forward PDEs, as in our target problem of global seismic inversion.

We have introduced a method that exploits the local structure of the posterior pdf—namely the Hessian matrix of the negative log posterior, which represents the local covariance—to overcome the curse of dimensionality associated with sampling high-dimensional distributions. Unfortunately, straight-

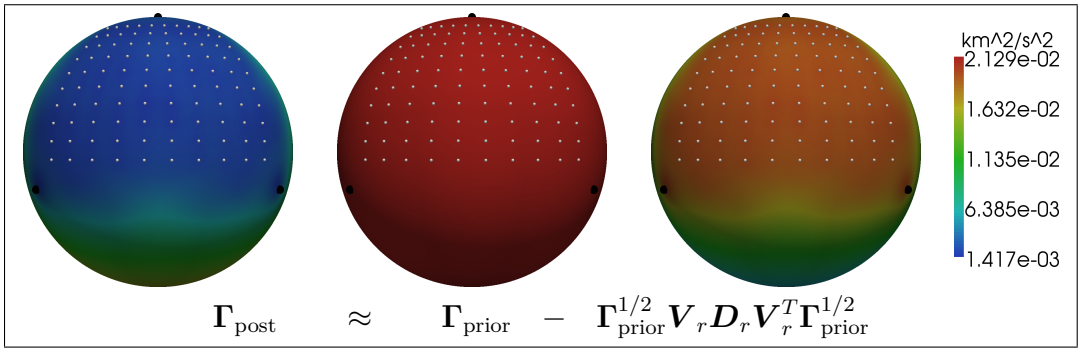

Figure 3.7: The left image depicts the pointwise posterior variance field, which is represented as the difference between the original prior variance field (middle), and the reduction in variance due to data (right; see also Figure 3.8). All variance fields are displayed at a depth of 67km.

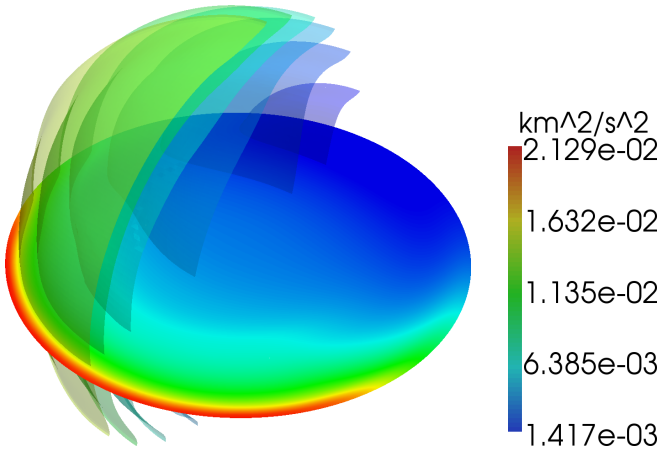

Figure 3.8: Data-induced reduction in variance inside the earth. The reduction is shown on a slice through the equator, as well as on isosurfaces in the left hemisphere (compare with Figure 3.7, which shows reduction on earth's surface). As can be seen, the reduction in variance is greatest on the surface.

forward computation of the dense Hessian is prohibitive, requiring as many forward-like solves as there are uncertain parameters. However, the data are typically informative about a low dimensional subspace of the parameter space—that is, the Hessian is sparse with respect to some basis. We have exploited this fact to construct a low rank approximation of the Hessian and its inverse using a matrix-free parallel randomized subspace-detecting algorithm. Overall, our method requires a dimension-independent number of forward PDE solves to approximate the local covariance. Uncertainty quantification for the inverse problem thus reduces to solving a fixed number of forward and adjoint PDEs (which resemble the original forward problem), independent of the problem dimension. The entire process is thus scalable with respect to the forward problem dimension, uncertain parameter dimension, observational data dimension, and number of processor cores. We applied this method to the Bayesian solution of an inverse problem in 3D global seismic wave propagation with one million inversion parameters, for which we observe 3 orders of magnitude dimension reduction, which makes UQ tractable. This is by far the largest UQ problem that has been solved with such a complex governing PDE model.

3.9 Acknowledgments

Support for this work was provided by: the U.S. Air Force Office of Scientific Research (AFOSR) Computational Mathematics program under award number FA9550-09-1-0608; the U.S. Department of Energy Office of

Science (DOE-SC), Advanced Scientific Computing Research (ASCR), Scientific Discovery through Advanced Computing (SciDAC) program, under award numbers DE-FC02-11ER26052 and DE-FG02-09ER25914, and the Multiscale Mathematics and Optimization for Complex Systems program under award number DE-FG02-08ER25860; the U.S. DOE National Nuclear Security Administration, Predictive Simulation Academic Alliance Program (PSAAP), under award number DE-FC52-08NA28615; and the U.S. National Science Foundation (NSF) Cyber-enabled Discovery and Innovation (CDI) program under awards CMS-1028889 and OPP-0941678, and the Collaborations in Mathematical Geosciences (CMG) program under award DMS-0724746. Computing time on the Cray XK6 supercomputer (Jaguar) was provided by the Oak Ridge Leadership Computing Facility at Oak Ridge National Laboratory, which is supported by the Office of Science of the Department of Energy under Contract DE-AC05-00OR22725. Computing time on the Texas Advanced Computing Center's Lonestar 4 supercomputer was provided by an allocation from TACC.

Chapter 4

Optimal low-rank approximations of Bayesian linear inverse problems

The content of this chapter is based on a submitted manuscript which is currently under review[1], and is joint work with Alessio Spantini, Antti Solonen, Tingang Cui, Luis Tenorio, and Youssef Marzouk. This is work that emerged from my early collaborations with Youssef on dimension reduction for Bayesian inverse problems. We observed empirically that the low-rank approximation we used performed very well, and I made many attempts to prove that it was optimal. I did not succeed in this, but uncovered a sufficient quantity of negative results to lead us to look at alternative metrics, including the Förstner metric discussed in this chapter. Alessio, working closely with Luis, deserves the major credit for the development of the proofs for optimality of the covariance approximation that were ultimately successful. Alessio and Antti later collaborated for the optimality results for the mean estimator. The numerical experiments for the tomography problem and randomized Hessians and priors are due to Antti, and Tiangang contributed the numerical

[1] A. Spantini, A. Solonen, T. Cui, J. Martin, L. Tenorio, and Y. Marzouk. Optimal low-rank approximations of Bayesian linear inverse problems. *Submitted*, 2014. arXiv preprint arXiv:1407.3463. http://arxiv.org/abs/1407.3463

experiments for the diffusion example.

Abstract

In the Bayesian approach to inverse problems, data are often informative, relative to the prior, only on a low-dimensional subspace of the parameter space. Significant computational savings can be achieved by using this subspace to characterize and approximate the posterior distribution of the parameters. We first investigate approximation of the posterior covariance matrix as a low-rank update of the prior covariance matrix. We prove optimality of a particular update, based on the leading eigendirections of the matrix pencil defined by the Hessian of the log-likelihood and the prior precision, for a broad class of loss functions. This class includes the Förstner metric for symmetric positive definite matrices, as well as the Kullback-Leibler divergence and the Hellinger distance between the associated distributions. We also propose two fast approximations of the posterior mean and prove their optimality with respect to a weighted Bayes risk under squared-error loss. These approximations are particularly useful when repeated posterior mean evaluations are required for multiple data sets. We demonstrate our theoretical results with several numerical examples, including high-dimensional X-ray tomography and an inverse heat conduction problem. In both of these examples, the intrinsic low-dimensional structure of the inference problem can be exploited while producing results that are essentially indistinguishable from solutions computed in the full space.

4.1 Introduction

In the Bayesian approach to inverse problems, the parameters of interest are treated as random variables, endowed with a prior probability distribution that encodes information available before any data are observed. Observations are modeled by their joint probability distribution conditioned on the parameters of interest, which defines the likelihood function and incorporates the forward model and a stochastic description of measurement or model errors. The prior and likelihood then combine to yield a probability distribution for the parameters conditioned on the observations, i.e., the posterior distribution. While this formulation is quite general, essential features of inverse problems bring additional structure to the Bayesian update. The prior distribution often encodes some kind of smoothness or correlation among the inversion parameters; observations typically are finite, few in number, and corrupted by noise; and the observations are indirect, related to the inversion parameters by the action of a forward operator that destroys some information. A key consequence of these features is that the data may be informative, relative to the prior, only on a *low-dimensional subspace* of the entire parameter space. Identifying and exploiting this subspace—to design approximations of the posterior distribution and related Bayes estimators—can lead to substantial computational savings.

In this paper we investigate approximation methods for finite-dimensional Bayesian linear inverse problems with Gaussian measurement and prior distributions. We characterize approximations of the posterior distribution that

are structure-exploiting and that are *optimal* in a sense to be defined below. Since the posterior distribution is Gaussian, it is completely determined by its mean and covariance. We therefore focus on approximations of these posterior characteristics. Optimal approximations will reduce computation and storage requirements for high-dimensional inverse problems, and will also enable fast computation of the posterior mean in a many-query setting.

We consider approximations of the posterior covariance matrix in the form of low-rank negative updates of the prior covariance matrix. This class of approximations exploits the structure of the prior-to-posterior update, and also arises naturally in Kalman filtering techniques (e.g., [12, 13, 185]); the challenge is to find an *optimal* update within this class, and to define in what sense it is optimal. We will argue that a suitable loss function with which to define optimality is the Förstner metric [75] for symmetric positive definite matrices, and will show that this metric generalizes to a broader class of loss functions that emphasize relative differences in covariance. We will derive the optimal low-rank update for this entire class of loss functions. In particular, we will show that the prior covariance matrix should be updated along the leading generalized eigenvectors of the pencil $(H, \Gamma_{\text{pr}}^{-1})$ defined by the Hessian of the log-likelihood and the prior precision matrix. If we assume exact knowledge of the posterior mean, then our results extend to optimality statements between distributions (e.g., optimality in Kullback-Leibler divergence and in Hellinger distance). The form of this low-rank update of the prior is not new [26, 33, 73, 143], but previous work has not shown whether—and if so, in exactly what

sense—it yields optimal approximations of the posterior. A key contribution of this paper is to establish and explain this optimality.

Properties of the generalized eigenpairs of $(H, \Gamma_{\text{pr}}^{-1})$ and related matrix pencils have been studied previously in the literature, especially in the context of classical regularization techniques for linear inverse problems[2] [62, 100, 101, 159, 197]. The joint action of the log-likelihood Hessian and the prior precision matrix has also been used in related regularization methods [38, 40, 41, 113]. However, these efforts have not been concerned with the posterior covariance matrix or with its optimal approximation, since this matrix is a property of the Bayesian approach to inversion.

One often justifies the assumption that the posterior mean is exactly known by arguing that it can easily be computed as the solution of a regularized least-squares problem [5, 17, 108, 148, 160]; indeed, evaluation of the posterior mean to machine precision is now feasible even for million-dimensional parameter spaces [26]. If, however, one needs multiple evaluations of the posterior mean for different realizations of the data (e.g., in an online inference context), then solving a linear system to determine the posterior mean may not be the most efficient strategy. A second goal of our paper is to address this problem. We will propose two computationally efficient approximations of the posterior mean based on: *(i)* evaluating a low-rank affine function of

[2] In the framework of Tikhonov regularization [194], the regularized estimate coincides with the posterior mean of the Bayesian linear model we consider here, provided that the prior covariance matrix is chosen appropriately.

the data; or *(ii)* using a low-rank update of the prior covariance matrix in the exact formula for the posterior mean. The optimal approximation in each case is defined as the minimizer of the Bayes risk for a squared-error loss defined by the posterior precision matrix. We provide explicit formulas for these optimal approximations and show that they can be computed by exploiting the optimal posterior covariance approximation described above. Thus, given a new set of data, computing an optimal approximation of the posterior mean becomes a computationally trivial task.

Low-rank approximations of the posterior mean that minimize the Bayes risk for squared-error loss under the standard Euclidean norm have been proposed in [49, 50]. Here, instead we develop analytical results for squared-error loss weighted by the posterior precision matrix. This choice of norm reflects the idea that approximation errors in directions of low posterior variance should be penalized more strongly than errors in high-variance directions, as we do not want the approximate posterior mean to fall outside the bulk of the posterior probability distribution. Remarkably, in this case, the optimal approximation only requires the leading eigenvectors and eigenvalues of a single eigenvalue problem. This is the same eigenvalue problem we solve to obtain an optimal approximation of the posterior covariance matrix, and thus we can efficiently obtain both approximations at the same time.

While the efficient solution of large-scale linear-Gaussian Bayesian inverse problems is of standalone interest [73], optimal approximations of Gaussian posteriors are also a building block for the solution of nonlinear Bayesian

inverse problems. For example, the stochastic Newton Markov chain Monte Carlo (MCMC) method [143] uses Gaussian proposals derived from local linearizations of a nonlinear forward model; the parameters of each Gaussian proposal are computed using the optimal approximations analyzed in this paper. To tackle even larger nonlinear inverse problems, [26] uses a Laplace approximation of the posterior distribution wherein the Hessian at the mode of the log-posterior density is itself approximated using the present approach. Similarly, approximations of local Gaussians can facilitate the construction of a nonstationary Gaussian process whose mean directly approximates the posterior density [32]. Alternatively, [57] combines data-informed directions derived from local linearizations of the forward model—a direct extension of the posterior covariance approximations described in the present work—to create a global *data-informed subspace*. A computationally efficient approximation of the posterior distribution is then obtained by restricting MCMC to this subspace and treating complementary directions analytically. Moving from the finite to the infinite-dimensional setting, the same global data-informed subspace is used to drive efficient dimension-independent posterior sampling for inverse problems in [56].

Earlier work on dimension reduction for Bayesian inverse problems used the Karhunen-Loève expansion of the prior distribution [133, 145] to describe the parameters of interest. To reduce dimension, this expansion is truncated; this step renders both the prior and posterior distributions singular—i.e., collapsed onto the prior mean—in the neglected directions. Avoiding large trunca-

tion errors then requires that the prior distribution impose significant smoothness on the parameters, so that the spectrum of the prior covariance kernel decays quickly. In practice, this requirement restricts the choice of priors. Moreover, this approach relies entirely on properties of the prior distribution and does not incorporate the influence of the forward operator or the observational errors. Alternatively, [136] constructs a reduced basis for the parameter space via greedy model-constrained sampling, but this approach can also fail to capture posterior variability in directions uninformed by the data. Both of these earlier approaches seek reduction in the overall description of the parameters. This notion differs fundamentally from the dimension reduction technique advocated in this paper, where low-dimensional structure is sought in the *change* from prior to posterior.

The rest of this paper is organized as follows. In Section 4.2 we introduce the posterior covariance approximation problem and derive the optimal prior-to-posterior update with respect to a broad class of loss functions. The structure of the optimal posterior covariance matrix approximation is examined in Section 4.3. Several interpretations are given in this section, including an equivalent reformulation of the covariance approximation problem as an optimal *projection* of the likelihood function onto a lower dimensional subspace. In Section 4.4 we characterize optimal approximations of the posterior mean. In Section 4.5 we provide several numerical examples. Section 4.6 offers concluding remarks. Appendix 4.7 collects proofs of many of the theorems stated throughout the paper, along with additional technical results.

4.2 Optimal approximation of the posterior covariance matrix

Consider the Bayesian linear model defined by a Gaussian likelihood and a Gaussian prior with a non-singular covariance matrix $\Gamma_{\text{pr}} \succ 0$ and, without loss of generality, zero mean:

$$y \mid x \sim \mathcal{N}(Gx, \Gamma_{\text{obs}}), \quad x \sim \mathcal{N}(0, \Gamma_{\text{pr}}). \tag{4.1}$$

Here x represents the parameters to be inferred, G is the linear forward operator, and y are the observations, with $\Gamma_{\text{obs}} \succ 0$. The statistical model (4.1) also follows from:

$$y = Gx + \varepsilon$$

where $\varepsilon \sim \mathcal{N}(0, \Gamma_{\text{obs}})$ is independent of x. It is easy to see that the posterior distribution is again Gaussian (see, e.g., [42]): $x \mid y \sim \mathcal{N}(\mu_{\text{pos}}(y), \Gamma_{\text{pos}})$, with mean and covariance matrix given by

$$\mu_{\text{pos}}(y) = \Gamma_{\text{pos}}\, G^{\top}\Gamma_{\text{obs}}^{-1}\, y \quad \text{and} \quad \Gamma_{\text{pos}} = \left(H + \Gamma_{\text{pr}}^{-1}\right)^{-1}, \tag{4.2}$$

where

$$H = G^{\top}\Gamma_{\text{obs}}^{-1}G \tag{4.3}$$

is the Hessian of the negative log-likelihood (i.e., the Fisher information matrix). Since the posterior is Gaussian, the posterior mean coincides with the posterior mode: $\mu_{\text{pos}}(y) = \arg\max_x \pi_{\text{pos}}(x; y)$, where π_{pos} is the posterior density. Note that the posterior covariance matrix does not depend on the data.

4.2.1 Defining the approximation class

We will seek an approximation, $\widehat{\Gamma}_{\text{pos}}$, of the posterior covariance matrix that is optimal in a class of matrices to be defined shortly. As we can see from (4.2), the posterior precision matrix Γ_{pos}^{-1} is a non-negative update of the prior precision matrix Γ_{pr}^{-1}: $\Gamma_{\text{pos}}^{-1} = \Gamma_{\text{pr}}^{-1} + ZZ^{\top}$, where $ZZ^{\top} = H$. Similarly, using Woodbury's identity we can write Γ_{pos} as a non-positive update of Γ_{pr}: $\Gamma_{\text{pos}} = \Gamma_{\text{pr}} - KK^{\top}$, where $KK^{\top} = \Gamma_{\text{pr}}\, G^{\top}\Gamma_{\text{y}}^{-1} G\, \Gamma_{\text{pr}}$ and $\Gamma_{\text{y}} = \Gamma_{\text{obs}} + G\, \Gamma_{\text{pr}}\, G^{\top}$ is the covariance matrix of the marginal distribution of y [117]. This update of Γ_{pr} is negative semidefinite because the data add information: the posterior variance in any direction is always smaller than the corresponding prior variance. Moreover, the update is usually low rank for exactly the reasons described in the introduction: there are directions in the parameter space along which the data are not very informative, relative to the prior. For instance, H might have a quickly decaying spectrum [29]. Note, however, that Γ_{pos} itself might *not* be low-rank. Low-rank structure, if any, lies in the update of Γ_{pr} that yields Γ_{pos}. Hence, a natural class of matrices for approximating Γ_{pos} is the set of negative semi-definite updates of Γ_{pr}, with a fixed maximum rank, that lead to positive definite matrices:

$$\mathcal{M}_r = \left\{ \Gamma_{\text{pr}} - KK^{\top} \succ 0 : \operatorname{rank}(K) \leq r \right\}. \tag{4.4}$$

This class of approximations of the posterior covariance matrix takes advantage of the structure of the prior-to-posterior update.

4.2.2 Loss functions

Optimality statements regarding the approximation of a covariance matrix require an appropriate notion of distance between symmetric positive definite (SPD) matrices. We shall use the metric introduced by Förstner and Moonen [75], which is derived from a canonical invariant metric on the cone of real SPD matrices and is defined as follows: the Förstner distance, $d_{\mathcal{F}}(A, B)$, between a pair of SPD matrices, A and B, is given by

$$d^2_{\mathcal{F}}(A, B) = \operatorname{tr}\left[\ln^2(\, A^{-1/2} B A^{-1/2}\,)\right] = \sum_i \ln^2(\sigma_i),$$

where (σ_i) is the sequence of generalized eigenvalues of the pencil (A, B). The Förstner metric satisfies the following important invariance properties:

$$d_{\mathcal{F}}(A,\, B) = d_{\mathcal{F}}(A^{-1},\, B^{-1}), \quad \text{and} \quad d_{\mathcal{F}}(A,\, B) = d_{\mathcal{F}}(MAM^\top,\, MBM^\top) \tag{4.5}$$

for any nonsingular matrix M. Moreover, $d_{\mathcal{F}}$ treats under- and over-approximations similarly in the sense that $d_{\mathcal{F}}(\Gamma_{\text{pos}},\, \alpha\widehat{\Gamma}_{\text{pos}}) \to \infty$ as $\alpha \to 0$ and as $\alpha \to \infty$.[3] Note that the metric induced by the Frobenius norm does not satisfy any of the aforementioned invariance properties. In addition, it penalizes under- and over-estimation differently.

We will show that our posterior covariance matrix approximation is optimal not only in terms of the Förstner metric, but also in terms of the

[3]This behavior is shared by Stein's loss function, which has been proposed to assess estimates of a covariance matrix [116]. Stein's loss function is just the Kullback-Leibler distance between two Gaussian distributions with the same mean (see (4.56)), but it is not a metric for SPD matrices.

following more general class, $\mathcal{L}$, of loss functions for SPD matrices.

Definition 1 (Loss functions)**.** *The class $\mathcal{L}$ is defined as the collection of functions of the form*

$$L(A, B) = \sum_{i=1}^{n} f(\sigma_i), \tag{4.6}$$

where A and B are SPD matrices, (σ_i) are the generalized eigenvalues of the pencil (A, B), and

$$f \in \mathcal{U} = \{g \in \mathcal{C}^1(\mathbb{R}_+) : g'(x)(1-x) < 0 \text{ for } x \neq 1, \text{ and } \lim_{x\to\infty} g(x) = \infty\}. \tag{4.7}$$

Elements of $\mathcal{U}$ are differentiable real-valued functions defined on the positive axis that decrease on $x < 1$, increase on $x > 1$, and tend to infinity as $x \to \infty$. The squared Förstner metric belongs to the class of loss functions defined by (4.6), whereas the distance induced by the Frobenius norm does not.

Lemma 1, whose proof can be found in Appendix 4.7, justifies the importance of the class $\mathcal{L}$. In particular, it shows that optimality of the covariance matrix approximation with respect to any loss function in $\mathcal{L}$ leads to an optimal approximation of the posterior distribution using a Gaussian (with the same mean) in terms of other familiar criteria used to compare probability measures, such as the Hellinger distance and the Kullback-Leibler (K-L) divergence [161]. More precisely, we have the following result:

Lemma 1 (Equivalence of approximations)**.** *If $L \in \mathcal{L}$, then a matrix $\widehat{\Gamma}_{\text{pos}} \in \mathcal{M}_r$ minimizes the Hellinger distance and the K-L divergence between $\mathcal{N}(\mu_{\text{pos}}(y), \Gamma_{\text{pos}})$ and the approximation $\mathcal{N}(\mu_{\text{pos}}(y), \widehat{\Gamma}_{\text{pos}})$ iff it minimizes $L(\Gamma_{\text{pos}}, \widehat{\Gamma}_{\text{pos}})$.*

Remark 1. We note that neither the Hellinger distance nor the K-L divergence between the distributions $\mathcal{N}(\mu_{\text{pos}}(y), \Gamma_{\text{pos}})$ and $\mathcal{N}(\mu_{\text{pos}}(y), \widehat{\Gamma}_{\text{pos}})$ depends on the data y. Optimality in distribution does not necessarily hold when the posterior means are different.

4.2.3 Optimality results

We are now in a position to state one of the main results of the paper. For a proof see Appendix 4.7.

Theorem 1 (Optimal posterior covariance approximation)**.** *Let $(\delta_i^2, \widehat{w}_i)$ be the generalized eigenvalue-eigenvector pairs of the pencil:*

$$(H, \Gamma_{\text{pr}}^{-1}), \tag{4.8}$$

with the ordering $\delta_i^2 \geq \delta_{i+1}^2$, and $H = G^\top \Gamma_{\text{obs}}^{-1} G$ as in (4.3). Let L be a loss function in the class $\mathcal{L}$ defined in (4.6). Then:

(i) A minimizer, $\widehat{\Gamma}_{\text{pos}}$, of the loss L between Γ_{pos} and an element of $\mathcal{M}_r$ is given by:

$$\widehat{\Gamma}_{\text{pos}} = \Gamma_{\text{pr}} - KK^\top, \quad KK^\top = \sum_{i=1}^{r} \delta_i^2 \left(1 + \delta_i^2\right)^{-1} \widehat{w}_i \widehat{w}_i^\top. \tag{4.9}$$

The corresponding minimum loss is given by:

$$L(\widehat{\Gamma}_{\text{pos}}, \Gamma_{\text{pos}}) = f\,(1)\, r + \sum_{i>r} f(\,1/(1+\delta_i^2)\,). \tag{4.10}$$

(ii) The minimizer (4.9) is unique if the first r eigenvalues of $(H, \Gamma_{\text{pr}}^{-1})$ are distinct.

Theorem 1 provides a way to construct the best approximation of Γ_{pos} by matrices in $\mathcal{M}_r$: it is just a matter of computing the eigenpairs corresponding to the decreasing sequence of eigenvalues of the pencil $(H, \Gamma_{\text{pr}}^{-1})$ until a stopping criterion is satisfied. This criterion can be based on the minimum loss (4.10). Notice that the minimum loss is a function of the generalized eigenvalues $(\delta_i^2)_{i \geq r}$ that have not been computed. This is quite common in numerical linear algebra (e.g., error in the truncated SVD [65, 89]). However, since the eigenvalues (δ_i^2) are computed in a decreasing order, the minimum loss can be easily bounded.

The generalized eigenvectors $\widehat{w}_i$ are orthogonal with respect to the inner product induced by the prior precision matrix, and they maximize the Rayleigh ratio,

$$\widehat{\mathcal{R}}(z) = \frac{z^\top H z}{z^\top \Gamma_{\text{pr}}^{-1} z},$$

over subspaces of the form $\widehat{\mathcal{W}}_i = \text{span}^\perp(\widehat{w}_j)_{j<i}$. Intuitively, the vectors $\widehat{w}_i$ associated with generalized eigenvalues greater than one correspond to directions in the parameter space (or subspaces thereof) where the curvature of the log-posterior density is constrained more by the log-likelihood than by the prior.

4.2.4 Computing eigenpairs of $(H, \Gamma_{\text{pr}}^{-1})$

If a square root factorization of the prior covariance matrix $\Gamma_{\text{pr}} = S_{\text{pr}} S_{\text{pr}}^\top$ is available, then the Hermitian generalized eigenvalue problem can be reduced to a standard one: find the eigenpairs (δ_i^2, w_i) of $S_{\text{pr}}^\top H S_{\text{pr}}$, and transform the

resulting eigenvectors according to $w_i \mapsto S_{\text{pr}} w_i$ [15, Section 5.2]. An analogous transformation is also possible when a square root factorization of Γ_{pr}^{-1} is available. Notice that only the actions of S_{pr} and $S_{\text{pr}}^\top$ on a vector are required. For instance, evaluating the action of S_{pr} might involve the solution of an elliptic PDE [137]. There are numerous examples of priors for which a decomposition $\Gamma_{\text{pr}} = S_{\text{pr}} S_{\text{pr}}^\top$ is readily available, e.g., [60, 137, 189, 202, 204]. Either direct methods or, more often, matrix-free algorithms (e.g., Lanczos iteration) can be used to solve the standard Hermitian eigenvalue problem [15, Section 4]. Reference implementations of these algorithms are available in ARPACK [129]. If a square root factorization of Γ_{pr} is not available, but it is possible to solve linear systems with Γ_{pr}^{-1}, we can use a Lanczos method for generalized Hermitian eigenvalue problems [15, Section 5.5] where a Krylov basis orthogonal with respect to the inner product induced by Γ_{pr}^{-1} is maintained. Again, the ARPACK provides an efficient implementation of these solvers. When solving accurate linear systems with Γ_{pr}^{-1} is a daunting task, we refer the reader to alternative algorithms proposed in [184] and [90].

Remark 2. If a factorization $\Gamma_{\text{pr}} = S_{\text{pr}} S_{\text{pr}}^\top$ is available, then it is straightforward to obtain an expression for a non-symmetric square root of the optimal approximation of Γ_{pos} (4.9) as in [33]:

$$\widehat{S}_{\text{pos}} = S_{\text{pr}} \left(\sum_{i=1}^{r} \left[\left(1 + \delta_i^2 \right)^{-1/2} - 1 \right] w_i w_i^\top + I \right) \tag{4.11}$$

such that $\widehat{\Gamma}_{\text{pos}} = \widehat{S}_{\text{pos}} \widehat{S}_{\text{pos}}^\top$ and $w_i = S_{\text{pr}}^{-1} \widehat{w}_i$. This expression can be used to efficiently sample from the approximate posterior distribution $\mathcal{N}(\mu_{\text{pos}}(y), \widehat{\Gamma}_{\text{pos}})$.

4.3 Properties of the optimal covariance approximation

Now we discuss several implications of the optimal approximation of Γ_{pos} introduced in the previous section. We start by describing the relationship between this approximation and the directions of greatest relative reduction of prior variance. Then we interpret the covariance approximation as the result of projecting the likelihood function onto a "data-informed" subspace. Finally, we contrast the present approach with several other approximation strategies: using the Frobenius norm as a loss function for the covariance matrix approximation, or developing low-rank approximations based on prior or Hessian information alone. We conclude by drawing the connections with the BFGS Kalman filter update.

4.3.1 Interpretation of the eigendirections

Thanks to the particular structure of loss functions in $\mathcal{L}$, the problem of approximating Γ_{pos} is equivalent to that of approximating Γ_{pos}^{-1}. Yet the form of the optimal approximation of Γ_{pos}^{-1} is important, as it explicitly describes the directions that control the ratio of posterior to prior variance. The following corollary to Theorem 1 characterizes these directions. The proof is in Appendix 4.7.

Corollary 1 (Optimal posterior precision approximation)**.** *Let* $(\delta_i^2, \widehat{w}_i)$ *and* $L \in \mathcal{L}$ *be defined as in Theorem 1. Then:*

(i) A minimizer of $L(B, \Gamma_{\text{pos}}^{-1})$ *for*

$$B \in \mathcal{M}_r^{-1} := \left\{ \Gamma_{\text{pr}}^{-1} + JJ^\top : \operatorname{rank}(J) \le r \right\} \tag{4.12}$$

is given by

$$\widehat{\Gamma}_{\text{pos}}^{-1} = \Gamma_{\text{pr}}^{-1} + UU^\top, \quad UU^\top = \sum_{i=1}^{r} \delta_i^2 \widetilde{w}_i \widetilde{w}_i^\top, \quad \widetilde{w}_i = \Gamma_{\text{pr}}^{-1} \widehat{w}_i. \tag{4.13}$$

The minimizer (4.13) *is unique if the first* r *eigenvalues of* $(H, \Gamma_{\text{pr}}^{-1})$ *are distinct.*

(ii) The optimal posterior precision matrix (4.13) *is precisely the inverse of the optimal posterior covariance matrix* (4.9)*.*

(iii) The vectors $\widetilde{w}_i$ *are generalized eigenvectors of the pencil* $(\Gamma_{\text{pos}}, \Gamma_{\text{pr}})$*:*

$$\Gamma_{\text{pos}}\, \widetilde{w}_i = \frac{1}{1 + \delta_i^2}\, \Gamma_{\text{pr}}\, \widetilde{w}_i. \tag{4.14}$$

Note that the definition of the class $\mathcal{M}_r^{-1}$ is analogous to that of $\mathcal{M}_r$. Indeed, Lemma 3 in Appendix 4.7 defines a bijection between these two classes.

The vectors $\widetilde{w}_i = \Gamma_{\text{pr}}^{-1} \widehat{w}_i$ are orthogonal with respect to the inner product defined by Γ_{pr}. By (4.14), we also know that $\widetilde{w}_i$ minimizes the generalized Rayleigh quotient,

$$\mathcal{R}(z) = \frac{z^\top \Gamma_{\text{pos}} z}{z^\top \Gamma_{\text{pr}}\, z} = \frac{\mathbb{V}\text{ar}(z^\top x \mid y)}{\mathbb{V}\text{ar}(z^\top x)}, \tag{4.15}$$

over subspaces of the form $\widetilde{\mathcal{W}}_i = \operatorname{span}^{\perp}(\widetilde{w}_j)_{j<i}$. This Rayleigh quotient is precisely the *ratio of posterior to prior variance* along a particular direction,

z, in the parameter space. The smallest values that $\mathcal{R}$ can take over the subspaces $\widetilde{\mathcal{W}}_i$ are exactly the smallest generalized eigenvalues of $(\Gamma_{\text{pos}}, \Gamma_{\text{pr}})$. In particular, the data are most informative along the first r eigenvectors $\widetilde{w}_i$ and, since

$$\mathcal{R}(\widetilde{w}_i) = \frac{\mathbb{V}\text{ar}(\widetilde{w}_i^\top x \mid y)}{\mathbb{V}\text{ar}(\widetilde{w}_i^\top x)} = \frac{1}{1+\delta_i^2}, \tag{4.16}$$

the posterior variance is smaller than the prior variance by a factor of $(1+\delta_i^2)^{-1}$. In the span of the other eigenvectors, $(\widetilde{w}_i)_{i>r}$, the data are not as informative. Hence, $(\widetilde{w}_i)$ are the directions along which the ratio of posterior to prior variance is minimized. Furthermore, a simple computation shows that these directions also maximize the relative difference between prior and posterior variance normalized by the prior variance. Indeed, if the directions $(\widetilde{w}_i)$ minimize (4.15) then they must also maximize $1 - \mathcal{R}(z)$, leading to:

$$1 - \mathcal{R}(\widetilde{w}_i) = \frac{\mathbb{V}\text{ar}(\widetilde{w}_i^\top x) - \mathbb{V}\text{ar}(\widetilde{w}_i^\top x \mid y)}{\mathbb{V}\text{ar}(\widetilde{w}_i^\top x)} = \frac{\delta_i^2}{1+\delta_i^2}. \tag{4.17}$$

4.3.2 Optimal projector

Since the data are most informative on a subspace of the parameter space, it should be possible to reduce the *effective dimension* of the inference problem in a manner that is consistent with the posterior approximation. This is essentially the content of the following corollary, which follows by a simple computation.

Corollary 2 (Optimal projector)**.** *Let $\widehat{\Gamma}_{\text{pos}}$ and the vectors $(\widehat{w}_i, \widetilde{w}_i)$ be defined as in Theorems 1 and 1. Consider the reduced forward operator $\widehat{G}_r = G \circ P_r$,*

where P_r is the oblique projector (i.e., $P_r^2 = P_r$):

$$P_r = \sum_{i=1}^{r} \widehat{w}_i \widetilde{w}_i^\top. \tag{4.18}$$

Then $\widehat{\Gamma}_{\text{pos}}$ is precisely the posterior covariance matrix corresponding to the Bayesian linear model:

$$y \mid x \sim \mathcal{N}(\widehat{G}_r\, x,\, \Gamma_{\text{obs}}), \quad x \sim \mathcal{N}(0,\, \Gamma_{\text{pr}}). \tag{4.19}$$

The projected Gaussian linear model (4.19) reveals the intrinsic dimensionality of the inference problem. The introduction of the optimal projector (4.18) is also useful in the context of dimensionality reduction for nonlinear inverse problems. In this case a particularly simple and effective approximation of the posterior density $\pi_{\text{pos}}(x|y)$ is of the form $\widehat{\pi}_{\text{pos}}(x|y) \propto \pi(y; P_r\, x)\, \pi_{\text{pr}}(x)$, where π_{pr} is the prior density and $\pi(y; P_r\, x)$ is the density corresponding to the likelihood function with parameters constrained by the projector. The range of the projector can be determined by combining locally optimal data-informed subspaces from high-density regions in the support of the posterior distribution. This approximation is the subject of a related paper [57].

Returning to the linear inverse problem, notice also that the posterior mean of the projected model (4.19) might be used as an efficient approximation of the exact posterior mean. We will show in Section 4.4 that this posterior mean approximation in fact minimizes the Bayes risk for a weighted squared-error loss among all low-rank linear functions of the data.

4.3.3 Comparison with optimality in Frobenius norm

Thus far our optimality results for the approximation of Γ_{pos} have been restricted to the class of loss functions $\mathcal{L}$ given in Definition 1. However, it is also interesting to investigate optimality in the metric defined by the Frobenius norm. Given any two matrices A and B of the same size, the Frobenius distance between them is defined as $\|A - B\|$, where $\|\cdot\|$ is the Frobenius norm. Note that the Frobenius distance does not exploit the structure of the positive definite cone of symmetric matrices. The matrix $\widehat{\Gamma}_{\text{pos}} \in \mathcal{M}_r$ that minimizes the Frobenius distance from the exact posterior covariance matrix is given by:

$$\widehat{\Gamma}_{\text{pos}} = \Gamma_{\text{pr}} - KK^{\top}, \quad KK^{\top} = \sum_{i=1}^{r} \lambda_i \, u_i u_i^{\top}, \tag{4.20}$$

where (u_i) are the directions corresponding to the r largest eigenvalues of $\Gamma_{\text{pr}} - \Gamma_{\text{pos}}$. This result can be very different from the optimal approximation given in Theorem 1. In particular, the directions (u_i) are solutions of the eigenvalue problem

$$\Gamma_{\text{pr}} \, G^{\top} \Gamma_{\text{y}}^{-1} G \, \Gamma_{\text{pr}} \, u = \lambda u, \tag{4.21}$$

which maximize

$$u^{\top}(\Gamma_{\text{pr}} - \Gamma_{\text{pos}})u = \mathbb{V}\text{ar}(u^{\top}x) - \mathbb{V}\text{ar}(u^{\top}x \mid y). \tag{4.22}$$

That is, while optimality in the Förstner metric identifies directions that maximize the *relative* difference between prior and posterior variance, the Frobenius distance favors directions that maximize only the absolute value of this difference. There are many reasons to prefer the former. For instance, data might

be informative along directions of low prior variance (perhaps due to inadequacies in prior modeling); a covariance matrix approximation that is optimal in Frobenius distance may ignore updates in these directions entirely. Also, if parameters of interest (i.e., components of x) have differing units of measurement, relative variance reduction provides a unit-independent way of judging the quality of a posterior approximation; this notion follows naturally from the second invariance property of $d_{\mathcal{F}}$ in (4.5). From a computational perspective, solving the eigenvalue problem (4.21) is quite expensive compared to finding the generalized eigenpairs of the pencil $(H, \Gamma_{\text{pr}}^{-1})$. Finally, optimality in the Frobenius distance for an approximation of Γ_{pos} does not yield an optimality statement for the corresponding approximation of the posterior distribution, as shown in Lemma 1 for loss functions in $\mathcal{L}$.

4.3.4 Suboptimal posterior covariance approximations

4.3.4.1 Hessian-based and prior-based reduction schemes

The posterior approximation described by Theorem 1 uses both Hessian and prior information. It is instructive to consider approximations of the linear Bayesian inverse problem that rely only on one or the other; as we will illustrate numerically in Section 4.5.1, these approximations can be viewed as natural limiting cases of our approach. They are also closely related to previous efforts in dimensionality reduction that propose only Hessian-based [135] or prior-based [145] reductions. In contrast with these previous efforts, here we will consider versions of Hessian- and prior-based reduction that do not discard

prior information in the remaining directions. In other words, we will discuss posterior covariance approximations that remain in the form of (4.4)—i.e., updating the prior covariance only in r directions.

A Hessian-based reduction scheme updates Γ_{pr} in directions where the data have greatest influence in an absolute sense (i.e., not relative to the prior). This involves approximating the minus log-likelihood Hessian (4.3) with a low-rank decomposition as follows: let (s_i^2, v_i) be the eigenvalue-eigenvector pairs of H with the ordering $s_i^2 \geq s_{i+1}^2$. Then a best low-rank approximation of H in the Frobenius norm is given by:

$$H \approx \sum_{i=1}^{r} s_i^2 \, v_i v_i^\top = V_r S_r V_r^\top,$$

where v_i is the ith column of V_r and $S_r = \text{diag}\{s_1^2, \ldots, s_r^2\}$. Using Woodbury's identity we then obtain an approximation of Γ_{pos} as a low-rank negative semidefinite update of Γ_{pr}:

$$\Gamma_{\text{pos}} \approx \left(V_r S_r V_r^\top + \Gamma_{\text{pr}}^{-1}\right)^{-1} = \Gamma_{\text{pr}} - \Gamma_{\text{pr}} V_r \left(S_r^{-1} + V_r^\top \Gamma_{\text{pr}} V_r\right)^{-1} V_r^\top \Gamma_{\text{pr}}. \quad (4.23)$$

This approximation of the posterior covariance matrix belongs to the class $\mathcal{M}_r$. Thus, Hessian-based reduction is in general suboptimal when compared to the optimal approximation defined in Theorem 1. Note that an equivalent way to obtain (4.23) is to use a reduced forward operator of the form $\widehat{G} = G \circ V_r V_r^\top$, which is the composition of the original forward operator with a projector onto the leading eigenspace of H. In general, the projector $P_r = V_r V_r^\top$ is different from the optimal projector defined in Corollary 2 and is thus suboptimal.

To achieve prior-based reduction, on the other hand, we restrict the Bayesian inference problem to directions in the parameter space that explain most of the prior variance. More precisely, we look for a rank-r orthogonal projector P_r that minimizes the mean squared-error defined as:

$$\mathcal{E}\left(P_r\right) = \mathbb{E}\left(\|x - P_r x\|^2\right), \tag{4.24}$$

where the expectation is taken over the prior distribution (assumed to have zero mean) and $\|\cdot\|$ is the standard Euclidean norm [114]. Let (t_i^2, u_i) be the eigenvalue-eigenvector pairs of Γ_{pr} ordered as $t_i^2 \geq t_{i+1}^2$. Then a minimizer of (4.24) is given by the projector P_r onto the leading eigenspace of Γ_{pr}: $P_r = \sum_{i=1}^{r} u_i u_i^\top = U_r U_r^\top$, where u_i is the ith column of U_r. The actual approximation of the linear inverse problem consists of using the projected forward operator, $\widehat{G} = G \circ U_r U_r^\top$. By direct comparison with the optimal projector defined in Corollary 2, we see that the prior-based reduction is suboptimal in general. Also in this case, the posterior covariance matrix with the projected Gaussian model can be written as a negative semidefinite update of Γ_{pr}:

$$\Gamma_{\text{pos}} \approx \Gamma_{\text{pr}} - U_r T_r [\,(\,U_r^\top H U_r\,)^{-1} + T_r\,]^{-1} T_r U_r^\top,$$

where $T_r = \text{diag}\{t_1^2, \ldots, t_r^2\}$. The double matrix inversion makes this low-rank update computationally challenging to implement. It is also not optimal, as shown in Theorem 1.

To summarize, both the Hessian and the prior-based dimensionality reduction techniques are suboptimal. These methods do not take into account

the interactions between the dominant directions of H and those of Γ_{pr}, and the relative importance of these quantities. This is a key feature of the optimal covariance approximation described in Theorem 1. Section 4.5.1 will illustrate conditions under which these interactions become essential.

4.3.4.2 Connections with the BFGS Kalman filter

The linear Bayesian inverse problem analyzed in this paper can be interpreted as the analysis step of a linear Bayesian filtering problem [69]. If the prior distribution corresponds to the forecast distribution at some time t, the posterior coincides with the so-called analysis distribution. In the linear case, with Gaussian process noise and observational errors, both of these distributions are Gaussian. The Kalman filter [118] is the Bayesian solution to this filtering problem. In [12] the authors propose a computationally feasible way to implement (and approximate) this solution in large-scale systems. The key observation is that when solving an SPD linear system of the form $Ax = b$ by means of BFGS or limited memory BFGS (L-BFGS [139]), one typically obtains an approximation of A^{-1} for free. This approximation can be written as a low-rank correction of an arbitrary positive definite initial approximation matrix A_0^{-1}. The matrix A_0^{-1} can be, for instance, the scaled identity. Notice that the approximation of A^{-1} given by L-BFGS is full rank and positive definite. This approximation is in principle convergent as the storage limit of L-BFGS increases [154]. An L-BFGS approximation of A is also possible [203].

There are many ways to exploit this property of the L-BFGS method.

In [12], for example, the posterior covariance is written as a low-rank update of the prior covariance matrix: $\Gamma_{\text{pos}} = \Gamma_{\text{pr}} - \Gamma_{\text{pr}} G^\top \Gamma_{\text{y}}^{-1} G \Gamma_{\text{pr}}$, where $\Gamma_{\text{y}} = \Gamma_{\text{obs}} + G\Gamma_{\text{pr}} G^\top$, and Γ_{y}^{-1} itself is approximated using the L-BFGS method. Since this approximation of Γ_{y} is full rank, however, this approach does not exploit potential low-dimensional structure of the inverse problem. Alternatively, one can obtain an L-BFGS approximation of Γ_{pos} when solving the linear system $\Gamma_{\text{pos}}^{-1} x = G^\top \Gamma_{\text{obs}}^{-1} y$ for the posterior mean $\mu_{\text{pos}}(y)$ [13]. If one uses the prior covariance matrix as an initial approximation matrix A_0^{-1}, then the resulting L-BFGS approximation of Γ_{pos} can be written as a low-rank update of Γ_{pr}. This approximation format is similar to the one discussed in [73] and advocated in this paper. However, the approach of [13] (or its ensemble version [185]) does not correspond to any known optimal approximation of the posterior covariance matrix, nor does it lead to any optimality statement between the corresponding probability distributions. This is an important contrast with the present approach, which we will revisit numerically in Section 4.5.1.

4.4 Optimal approximation of the posterior mean

In this section, we develop and characterize fast approximations of the posterior mean that can be used, for instance, to accelerate repeated inversion with multiple data sets. Note that we are not proposing alternatives to the efficient computation of the posterior mean for a single realization of the data. This task can already be accomplished with current state-of-the-art iterative solvers for regularized least-squares problems [5, 17, 108, 148, 160]. Instead, we

are interested in constructing *statistically optimal approximations* of the posterior mean as a linear function of the data. This function is data-independent, and computing it is more expensive than solving a single linear system. Subsequently using this approximation, however, is inexpensive as it does not involve the solution of any linear system. Our approach is therefore justified when the posterior mean must be evaluated for multiple instances of the data. This approach can thus be viewed as an offline–online strategy, where a more costly but data-independent offline calculation is followed by fast online evaluations.

4.4.1 Optimality results

For the Bayesian linear model defined in (4.1), the posterior mode is equal to the posterior mean, $\mu_{\text{pos}}(y) = \mathbb{E}(x|y)$, which is the minimizer of the Bayes risk for squared-error loss [128, 138]. We first review this idea and establish some basic notation. Let S be an SPD matrix and let

$$L(\delta(y), x) = (x - \delta(y))^\top S\,(x - \delta(y)) = \|x - \delta(y)\|_S^2$$

be the loss incurred by the estimator $\delta(y)$ of x. The Bayes risk $R\,(\delta(y), x)$ of $\delta(y)$ is defined as the average loss over the joint distribution of x and y [42,128]: $R(\delta(y), x) = \mathbb{E}\,(\,L(\delta(y), x)\,)$. Since

$$R(\delta(y), x) = \mathbb{E}\left(\|\,\delta(y) - \mu_{\text{pos}}(y)\,\|_S^2\right) + \mathbb{E}\left(\|\mu_{\text{pos}}(y) - x\|_S^2\right), \tag{4.25}$$

it follows that $\delta(y) = \mu_{\text{pos}}(y)$ minimizes the Bayes risk over all estimators of x.

To study approximations of $\mu_{\text{pos}}(y)$, we use the squared-error loss function defined by the Mahalanobis distance [48] induced by Γ_{pos}^{-1}: $L(\delta(y), x) = \|\delta(y) - x\|^2_{\Gamma_{\text{pos}}^{-1}}$. This loss function accounts for the geometry induced by the posterior measure on the parameter space, penalizing errors in the approximation of $\mu_{\text{pos}}(y)$ more strongly in directions of lower posterior variance.

Under the assumption of zero prior mean, $\mu_{\text{pos}}(y)$ is a linear function of the data. Hence we seek approximations of $\mu_{\text{pos}}(y)$ of the form Ay, where A is a matrix in a class to be defined. Our goal is to obtain fast posterior mean approximations that can be applied repeatedly to multiple realizations of y. We will thus consider two classes of approximation matrices:

$$\mathcal{A}_r := \{A \,:\, \text{rank}(A) \leq r\} \quad \text{and} \quad \widehat{\mathcal{A}}_r := \left\{A = (\Gamma_{\text{pr}} - B)\, G^\top \Gamma_{\text{obs}}^{-1} \,:\, \text{rank}(B) \leq r\right\}. \tag{4.26}$$

The class $\mathcal{A}_r$ consists of low-rank matrices; it is standard in the statistics literature [114]. The class $\widehat{\mathcal{A}}_r$, on the other hand, can be understood via comparison with (4.2); it simply replaces Γ_{pos} with a low-rank negative semidefinite update of Γ_{pr}. We shall henceforth use $\mathcal{A}$ to denote either of the two classes above.

Let $R_{\mathcal{A}}(Ay, x)$ be the Bayes risk of Ay subject to $A \in \mathcal{A}$. We may now restate our goal as: find $A^* \in \mathcal{A}$ that minimizes the Bayes risk $R_{\mathcal{A}}(Ay, x)$. That is, find

$$R_{\mathcal{A}}(A^* y, x) = \min_{A \in \mathcal{A}} \mathbb{E}(\, \|Ay - x\|^2_{\Gamma_{\text{pos}}^{-1}}\,). \tag{4.27}$$

The following two theorems show that for either class of approximation matrices, $\mathcal{A}_r$ or $\widehat{\mathcal{A}}_r$, this problem admits a particularly simple analytical solution

that exploits the structure of the optimal approximation of Γ_{pos}. The proofs of the theorems rely on a result by Friedland and Torokhti [77], and are given in Appendix 4.7. We also use the fact that $\mathbb{E}\left(\|\mu_{\text{pos}}(y) - x\|^2_{\Gamma^{-1}_{\text{pos}}}\right) = \ell$, where ℓ is the dimension of the parameter space.

Theorem 2. *Let $(\delta_i^2, \widehat{w}_i)$ be defined as in Theorem 1 and let $(\widehat{v}_i)$ be generalized eigenvectors of the pencil $(G\Gamma_{\text{pr}}G^\top, \Gamma_{\text{obs}})$ associated with a non-increasing sequence of eigenvalues. Then:*

(i) A solution of (4.27) for $A \in \mathcal{A}_r$ is given by:

$$A^* = \sum_{i=1}^{r} \frac{\delta_i}{1+\delta_i^2}\, \widehat{w}_i \widehat{v}_i^\top, \tag{4.28}$$

(ii) The corresponding minimum Bayes risk over $\mathcal{A}_r$ is given by:

$$R_{\mathcal{A}_r}(A^*y, x) = \mathbb{E}\left(\|A^*y - \mu_{\text{pos}}(y)\|^2_{\Gamma^{-1}_{\text{pos}}}\right) + \mathbb{E}\left(\|\mu_{\text{pos}}(y) - x\|^2_{\Gamma^{-1}_{\text{pos}}}\right) = \sum_{i>r} \delta_i^2 + \ell. \tag{4.29}$$

Notice that the rank-r posterior mean approximation given by Theorem 2 coincides with the posterior mean of the projected linear Gaussian model defined in (4.19). Applying this approximation to a new realization of the data then requires only a *low-rank* matrix-vector product, a computationally trivial task.

Remark 3. Equation (4.28) can be interpreted as the truncated GSVD solution of a Tikhonov regularized linear inverse problem [101] (with unit regularization parameter). Hence, Theorem 2 also describes a Bayesian property of the (frequentist) truncated GSVD estimator.

Remark 4. If factorizations of the form $\Gamma_{\text{pr}} = S_{\text{pr}} S_{\text{pr}}^\top$ and $\Gamma_{\text{obs}} = S_{\text{obs}} S_{\text{obs}}^\top$ are readily available, then we can characterize the triplets $(\delta_i, \widehat{w}_i, \widehat{v}_i)$ from a singular value decomposition, $S_{\text{obs}}^{-1} G S_{\text{pr}} = \sum_{i \geq 1} \delta_i v_i w_i^\top$, of the matrix $S_{\text{obs}}^{-1} G S_{\text{pr}}$ with the transformations $\widehat{w}_i = S_{\text{pr}} w_i$, $\widehat{v}_i = S_{\text{obs}}^{-\top} v_i$ and the ordering $\delta_i \geq \delta_{i+1}$. In particular, the approximate posterior mean can be written as:

$$\mu_{\text{pos}}^{(r)}(y) = S_{\text{pr}} (S_{\text{obs}}^{-1} G S_{\text{pr}})_r^{\text{Tikh}} S_{\text{obs}}^{-1} y \tag{4.30}$$

where $(S_{\text{obs}}^{-1} G S_{\text{pr}})_r^{\text{Tikh}}$ is the best rank r approximation to a Tikhonov regularized inverse.[4] That is, for any matrix A, $(A)_r$ is the best rank r approximation of A (e.g., computed via SVD), whereas $(A)^{\text{Tikh}} := (A^\top A + I)^{-1} A^\top$.

Theorem 3. *Let $\widehat{\Gamma}_{\text{pos}} \in \mathcal{M}_r$ be the optimal approximation of Γ_{pos} defined in Theorem 1. Then:*

(i) A solution of (4.27) for $A \in \widehat{\mathcal{A}}_r$ is given by:

$$\widehat{A}^* = \widehat{\Gamma}_{\text{pos}} \, G^\top \Gamma_{\text{obs}}^{-1}. \tag{4.31}$$

(ii) The corresponding minimum Bayes risk over $\widehat{\mathcal{A}}_r$ is given by:

$$R_{\widehat{\mathcal{A}}_r}(\widehat{A}^* y, x) = \mathbb{E}\left(\left\| \widehat{A}^* y - \mu_{\text{pos}}(y) \right\|_{\Gamma_{\text{pos}}^{-1}}^2 \right) + \mathbb{E}\left(\left\| \mu_{\text{pos}}(y) - x \right\|_{\Gamma_{\text{pos}}^{-1}}^2 \right) = \sum_{i > r} \delta_i^6 + \ell. \tag{4.32}$$

[4] With unit regularization parameter and identity regularization operator [102].

For the estimator described in Theorem 3, once the optimal approximation of Γ_{pos} is computed, the cost of approximating $\mu_{\text{pos}}(y)$ for a new realization of y is dominated by the adjoint and prior solves needed to apply $G^\top$ and Γ_{pr}, respectively. Combining the optimal approximations of $\mu_{\text{pos}}(y)$ and Γ_{pos} given by Theorems 3 and 1, respectively, yields a complete approximation of the Gaussian posterior distribution. This is precisely the approximation adopted by the stochastic Newton MCMC method [143] to describe the Gaussian proposal distribution obtained from a local linearization of the forward operator of a nonlinear Bayesian inverse problem. Our results support the algorithmic choice of [143] with precise optimality statements.

It is worth noting that the two optimal Bayes risks, (4.29) and (4.32), depend on the parameter r that defines the dimension of the corresponding approximation classes $\mathcal{A}_r$ and $\widehat{\mathcal{A}}_r$. In the former case, r is the rank of the optimal matrix that defines the approximation. In the latter case, r is the rank of a negative update of Γ_{pr} that yields the posterior covariance matrix approximation. We shall thus refer to the estimator given by Theorem 2 as the low-rank approximation and to the estimator given by Theorem 3 as the low-rank *update* approximation. In both cases, we shall refer to r as the order of the approximation. A posterior mean approximation of order r will be called *under-resolved* if more than r generalized eigenvalues of the pencil $(H, \Gamma_{\text{pr}}^{-1})$ are greater than one. If this is the case, then using the low-rank update approximation is not appropriate because the associated Bayes risk includes high-order powers of eigenvalues of $(H, \Gamma_{\text{pr}}^{-1})$ that are greater than one. Thus,

under-resolved approximations tend to be more accurate when using the low-rank approximation. As we will show in Section 4.5, this estimator is also less expensive to compute than its counterpart in Theorem 3. If, on the other hand, fewer than r eigenvalues of $(H, \Gamma_{\text{pr}}^{-1})$ are greater than one, then the optimal low-rank *update* estimator will have better performance than the optimal low-rank estimator in the following statistical sense:

$$0 < R_{\mathcal{A}_r}(A^*y, x) - R_{\widehat{\mathcal{A}}_r}(\widehat{A}^*y, x) = \sum_{i>r} \delta_i^2 \left(1 + \delta_i^2\right) \left(1 - \delta_i^2\right).$$

4.4.2 Connection with "priorconditioners"

In this subsection, we draw connections between the low-rank approximation of the posterior mean given in Theorem 2 and the regularized solution of a discrete ill-posed inverse problem $y = Gx + \varepsilon$ (using the notation of this paper) as presented in [38, 41]. In [38, 41], the authors propose an early stopping regularization using iterative solvers preconditioned by prior statistical information on the parameter of interest, say $x \sim \mathcal{N}(0, \Gamma_{\text{pr}})$, and on the noise, say $\varepsilon \sim \mathcal{N}(0, \Gamma_{\text{obs}})$.[5] That is, if factorizations $\Gamma_{\text{pr}} = S_{\text{pr}} S_{\text{pr}}^\top$ and $\Gamma_{\text{obs}} = S_{\text{obs}} S_{\text{obs}}^\top$ are available, then [41] provides a solution $x = S_{\text{pr}}\, q$ to the inverse problem, where q comes from an early stopping regularization applied to the preconditioned linear system:

$$S_{\text{obs}}^{-1} G S_{\text{pr}} q = S_{\text{obs}}^{-1} y. \tag{4.33}$$

[5] It suffices to consider a Gaussian approximation to the distribution of x and ε

The iterative method of choice in this case is the CGLS algorithm [41, 99] (or GMRES for nonsymmetric square systems [39]) equipped with a proper stopping criterion (e.g., the discrepancy principle [117]). Although the approach of [41] is not exactly Bayesian, we can still use the optimality results of Theorem 2 to justify the good performance of this particular form of regularization. By a property of the CGLS algorithm, the rth iterate $x^r = S_{\text{pr}} q^r$ satisfies:

$$q^r = \arg \min_{q \in \mathcal{K}_r\left(\widehat{H}, \widehat{y}\right)} \|S_{\text{obs}}^{-1} y - S_{\text{obs}}^{-1} G S_{\text{pr}} q\|. \tag{4.34}$$

where $\mathcal{K}_r(\widehat{H}, \widehat{y})$ is the r–dimensional Krylov subspace associated with the matrix $\widehat{H} = S_{\text{pr}}^{\top} H S_{\text{pr}}$ and starting vector $\widehat{y} = S_{\text{pr}}^{\top} G^{\top} \Gamma_{\text{obs}}^{-1} y$. It was shown in [107] that the CGLS solution, at convergence, can be written as $x^* = S_{\text{pr}}(S_{\text{obs}}^{-1} G S_{\text{pr}})^{\dagger} S_{\text{obs}}^{-1} y$, where $(\,\cdot\,)^{\dagger}$ denotes the Moore-Penrose pseudoinverse [150, 164]. To highlight the differences between the CGLS solution and (4.30), we assume that $\mathcal{K}_r(\widehat{H}, y) \approx \text{ran}(W_r)$ for all r, where $W_r = [w_1 \,|\, \cdots \,|\, w_r]$ and $\widehat{H} = \sum_i \delta_i^2 w_i w_i^{\top}$ is an SVD of $\widehat{H}$. Notice that the condition $\mathcal{K}_r(\widehat{H}, y) \approx \text{ran}(W_r)$ is usually quite reasonable for moderate values of r; this practical observation is at the heart of the Lanczos iteration for symmetric eigenvalue problems [124]. With simple algebraic manipulations we conclude that:

$$x^r \approx S_{\text{pr}}(S_{\text{obs}}^{-1} G S_{\text{pr}})_r^{\dagger}\, S_{\text{obs}}^{-1} y. \tag{4.35}$$

Recall from (4.30) that the optimal rank–r approximation of the posterior mean defined in Theorem 2 can be written as:

$$\mu_{\text{pos}}^{(r)}(y) = S_{\text{pr}}(S_{\text{obs}}^{-1} G S_{\text{pr}})_r^{\text{Tikh}} S_{\text{obs}}^{-1} y. \tag{4.36}$$

The only difference between (4.35) and (4.36) is the use of a Tikhonov-regularized inverse in (4.36) as opposed to a Moore-Penrose pseudoinverse. If $S_{\text{obs}}^{-1}GS_{\text{pr}} = \sum_{i\geq 1}\delta_i v_i w_i^\top$ is a reduced SVD of the matrix $S_{\text{obs}}^{-1}GS_{\text{pr}}$, then:

$$(S_{\text{obs}}^{-1}GS_{\text{pr}})_r^\dagger = \sum_{i\leq r}\frac{1}{\delta_i}w_i v_i^\top, \qquad (S_{\text{obs}}^{-1}GS_{\text{pr}})_r^{\text{Tikh}} = \sum_{i\leq r}\frac{\delta_i}{1+\delta_i^2}w_i v_i^\top. \tag{4.37}$$

These two matrices are *nearly* identical for values of r corresponding to δ_r^2 greater than one[6] (assuming the ordering $\delta_i^2 \geq \delta_{i+1}^2$). Beyond this regime, it might be convenient to stop the CGLS solver to obtain (4.35) (i.e., early stopping regularization). The similarity of these expressions is quite remarkable since (4.36) was derived as the minimizer of the optimization problem (4.27) with $\mathcal{A} = \mathcal{A}_r$. This informal argument may explain why *priorconditioners* perform so well in applications [40, 113]. Yet we remark that the goals of Theorem 2 and of [41] remain rather different; [41] is concerned with preconditioning techniques for early stopping regularization of ill-posed inverse problems, whereas Theorem 2 is concerned with statistically optimal approximations of the posterior mean in the Bayesian framework.

4.5 Numerical examples

Now we provide several numerical examples to illustrate the theory developed in the preceding sections. We start with a synthetic example to

[6] In Section 4.5 we show that by the time we start including generalized eigenvalues $\delta_i^2 \approx 1$ in (4.28), the approximation of the posterior mean is usually already satisfactory. Intuitively, this means that all the directions in parameter space where the data are more informative than the prior have been considered.

Algorithm 3 Optimal *low-rank* approximation of the posterior mean

INPUT: forward and adjoint models G, $G^\top$; prior and noise precisions Γ_{pr}^{-1}, Γ_{obs}^{-1}; approximation order $r \in \mathbb{N}$

OUTPUT: approximate posterior mean $\mu_{\text{pos}}^{(r)}(y)$

1: Find the r leading generalized eigenvalue-eigenvector pairs $(\delta_i^2, \widehat{w}_i)$ of the pencil $(G^\top \Gamma_{\text{obs}}^{-1} G, \Gamma_{\text{pr}}^{-1})$

2: Find the r leading generalized eigenvector pairs $(\widehat{v}_i)$ of the pencil $(G\,\Gamma_{\text{pr}} G^\top, \Gamma_{\text{obs}})$

3: For each new realization of the data y, compute $\mu_{\text{pos}}^{(r)}(y) = \sum_{i=1}^{r} \delta_i (1 + \delta_i^2)^{-1} \widehat{w}_i \widehat{v}_i^\top\, y$.

Algorithm 4 Optimal *low-rank update* approximation of the posterior mean

INPUT: forward and adjoint models G, $G^\top$; prior and noise precisions Γ_{pr}^{-1}, Γ_{obs}^{-1}; approximation order $r \in \mathbb{N}$

OUTPUT: approximate posterior mean $\widehat{\mu}_{\text{pos}}^{(r)}(y)$

1: Obtain $\widehat{\Gamma}_{\text{pos}}$ as described in Theorem 1.

2: For each new realization of the data y, compute $\widehat{\mu}_{\text{pos}}^{(r)}(y) = \widehat{\Gamma}_{\text{pos}}\, G^\top \Gamma_{\text{obs}}^{-1}\, y$.

demonstrate various posterior covariance matrix approximations, and continue with two more realistic linear inverse problems where we also study posterior mean approximations.

4.5.1 Example 1: Hessian and prior with controlled spectra

We begin by investigating the approximation of Γ_{pos} as a negative semidefinite update of Γ_{pr}. We compare the optimal approximation obtained in Theorem 1 with the Hessian-, prior-, and BFGS-based reduction schemes discussed in Section 4.3.4. The idea is to reveal differences between these approximations by exploring regimes where the data have differing impacts on the prior information. Since the directions defining the optimal update are

the generalized eigenvectors of the pencil $(H, \Gamma_{\text{pr}}^{-1})$, we shall also refer to this update as the *generalized* approximation.

To compare these approximation schemes, we start with a simple example employing diagonal Hessian and prior covariance matrices: $G = I$, $\Gamma_{\text{obs}} = \text{diag}\{\sigma_i^2\}$, and $\Gamma_{\text{pr}} = \text{diag}\{\lambda_i^2\}$. Since the forward operator G is the identity, this problem can (loosely) be thought of as denoising a signal x. In this case, $H = \Gamma_{\text{obs}}^{-1}$ and $\Gamma_{\text{pos}} = \text{diag}\{\lambda_i^2\sigma_i^2/(\sigma_i^2 + \lambda_i^2)\}$. The ratios of posterior to prior variance in the canonical directions (e_i) are

$$\frac{\mathbb{V}\text{ar}(e_i^\top x \mid y)}{\mathbb{V}\text{ar}(e_i^\top x)} = \frac{1}{1 + \lambda_i^2/\sigma_i^2}.$$

We note that if the observation variances σ_i^2 are constant, $\sigma_i = \sigma$, then the directions of greatest variance reduction are those corresponding to the largest prior variance. Hence the prior distribution alone determines the most informed directions, and the prior-based reduction is as effective as the generalized one. On the other hand, if the prior variances λ_i^2 are constant, $\lambda_i = \lambda$, but the σ_i vary, then the directions of highest variance reduction are those corresponding to the smallest noise variance. This time the noise distribution alone determines the most important directions, and Hessian-based reduction is as effective as the generalized one. In the case of more general spectra, the important directions depend on the ratios λ_i^2/σ_i^2 and thus one has to use the information provided by both the noise and prior distributions. This is done naturally by the generalized reduction.

We now generalize this simple case by moving to full matrices H and

Γ_{pr} with a variety of prescribed spectra. We assume that H and Γ_{pr} have SVDs of the form $H = U\Lambda U^\top$ and $\Gamma_{\text{pr}} = V\widetilde{\Lambda}V^\top$, where $\Lambda = \text{diag}\{\lambda_1, \ldots, \lambda_n\}$ and $\widetilde{\Lambda} = \text{diag}\{\widetilde{\lambda}_1, \ldots, \widetilde{\lambda}_n\}$ with

$$\lambda_k = \lambda_0/k^\alpha + \tau \quad \text{and} \quad \widetilde{\lambda}_k = \widetilde{\lambda}_0/k^{\widetilde{\alpha}} + \widetilde{\tau}.$$

To consider many different cases, the orthogonal matrices U and V are randomly and independently generated uniformly over the orthogonal group [187], leading to different realizations of H and Γ_{pr}. In particular, U and V are computed with a QR decomposition of a square matrix of independent standard Gaussian entries using a Gram-Schmidt orthogonalization. (In this case, the standard Householder reflections cannot be used.)

Before discussing the results of the first experiment, we explain our implementation of BFGS-based reduction. We ran the BFGS optimizer with a dummy quadratic optimization target $\mathcal{J}(x) = \frac{1}{2}\, x^\top \Gamma_{\text{pos}}^{-1} x$ and used Γ_{pr} as the initial approximation matrix for Γ_{pos}. Thus, the BFGS approximation of the posterior covariance matrix can be written as $\Gamma_{\text{pos}} = \Gamma_{\text{pr}} - KK^\top$ for some rank–r matrix K. The rank–r update is constructed by running the BFGS optimizer for r steps from random initial conditions as shown in [13]. Note that in order to obtain results for sufficiently high-rank updates, we use BFGS rather than L-BFGS in our numerical examples. While [12, 13] in principle employ L-BFGS, the results in these papers use a number of optimization steps roughly equal to the number of vectors stored in L-BFGS; our approach thus is comparable to [12, 13]. Nonetheless, some results for the highest-rank BFGS updates are

not plotted in Figures 4.1 and 4.2, as the optimizer converged so close to the optimum that taking further steps resulted in numerical instabilities.

Figure 4.1 summarizes the results of the first experiment. The top row shows the prescribed spectra of H^{-1} (red) and Γ_{pr} (blue). The parameters describing the eigenvalues of Γ_{pr} are fixed to $\widetilde{\lambda}_0 = 1$, $\widetilde{\alpha} = 2$, and $\widetilde{\tau} = 10^{-6}$. The corresponding parameters for H are given by $\lambda_0 = 500$ and $\tau = 10^{-6}$ with $\alpha = 0.345$ (left), $\alpha = 0.690$ (middle), and $\alpha = 1.724$ (right). Thus, moving from the leftmost column to the rightmost column, the data become increasingly less informative. The second row in the figure shows the Förstner distance between Γ_{pos} and its approximation, $\widehat{\Gamma}_{\text{pos}} = \Gamma_{\text{pr}} - KK^{\top}$, as a function of the rank of $KK^{\top}$ for 100 different realizations of H and Γ_{pr}. The third row shows, for each realization of (H, Γ_{pr}) and for each fixed rank of $KK^{\top}$, the difference between the Förstner distance obtained with a prior-, Hessian-, or BFGS-based dimensionality reduction technique and the minimum distance obtained with the generalized approximation. All of these differences are positive—a confirmation of Theorem 1. But Figure 4.1 also shows interesting patterns consistent with the observations made for the simple example above. When the spectrum of H is basically flat (left column), the directions along which the prior variance is most reduced are likely to be those corresponding to the largest prior variances, and thus a prior-based reduction is almost as effective as the generalized one (as seen in the bottom two rows on the left). As we move to the third column, eigenvalues of H^{-1} increase more quickly. The data provide significant information only on a lower-dimensional subspace of the parameter

space. In this case, it is crucial to combine this information with the directions in the parameter space along which the prior variance is the greatest. The generalized reduction technique successfully accomplishes this task, whereas the prior and Hessian reductions fail as they focus either on Γ_{pr} or H alone; the key is to combine the two. The BFGS update performs remarkably well across all three configurations of the Hessian spectrum, although it is clearly suboptimal compared to the generalized reduction.

In Figure 4.2 the situation is reversed and the results are symmetric to those of Figure 4.1. The spectrum of H (red) is now kept fixed with parameters $\lambda_0 = 500$, $\alpha = 1$, and $\tau = 10^{-9}$, while the spectrum of Γ_{pr} (blue) has parameters $\widetilde{\lambda}_0 = 1$ and $\widetilde{\tau} = 10^{-9}$ with decay rates increasing from left to right: $\widetilde{\alpha} = 0.552$ (left), $\widetilde{\alpha} = 1.103$ (middle), and $\widetilde{\alpha} = 2.759$ (right). In the first column, the spectrum of the prior is nearly flat. That is, the prior variance is almost equally spread along every direction in the parameter space. In this case, the eigenstructure of H determines the directions of greatest variance reduction, and the Hessian-based reduction is almost as effective as the generalized one. As we move towards the third column, the spectrum of Γ_{pr} decays more quickly. The prior variance is restricted to lower-dimensional subspaces of the parameter space. Mismatch between prior- and Hessian-dominated directions then leads to poor performance of both the prior- and Hessian-based reduction techniques. However, the generalized reduction performs well also in this more challenging case. The BFGS reduction is again empirically quite effective in most of the configurations that we consider. It is not always better

than the prior- or Hessian-based techniques when the update rank is low, or when the prior spectrum decays slowly; for example, Hessian-based reduction is more accurate than BFGS across all ranks in the first column of Figure 4.2. But when either the prior covariance or the Hessian have quickly decaying spectra, the BFGS approach performs almost as well as the generalized reduction. Though this approach remains suboptimal, its approximation properties bear further theoretical study.

4.5.2 Example 2: X-ray tomography

We consider a classical inverse problem of X-ray computed tomography (CT), where X-rays travel from sources to detectors through an object of interest. The intensities from multiple sources are measured at the detectors, the goal is to reconstruct the density of the object. In this framework, we investigate the performance of the optimal mean and covariance matrix approximations presented in Sections 4.2 and 4.4. This synthetic example is motivated by a real application: real-time X-ray imaging of logs that enter a saw mill for the purpose of automatic quality control. For instance, in the system commercialized by Bintec (www.bintec.fi), logs enter the X-ray system on fast-moving conveyer belt and quick reconstructions are needed. The imaging setting (e.g., X-ray source and detector locations) and the priors are fixed; only the data changes from one log cross-section to another. The basis for our posterior mean approximation can therefore be pre-computed, and repeated inversions can be carried out quickly with direct matrix formulas.

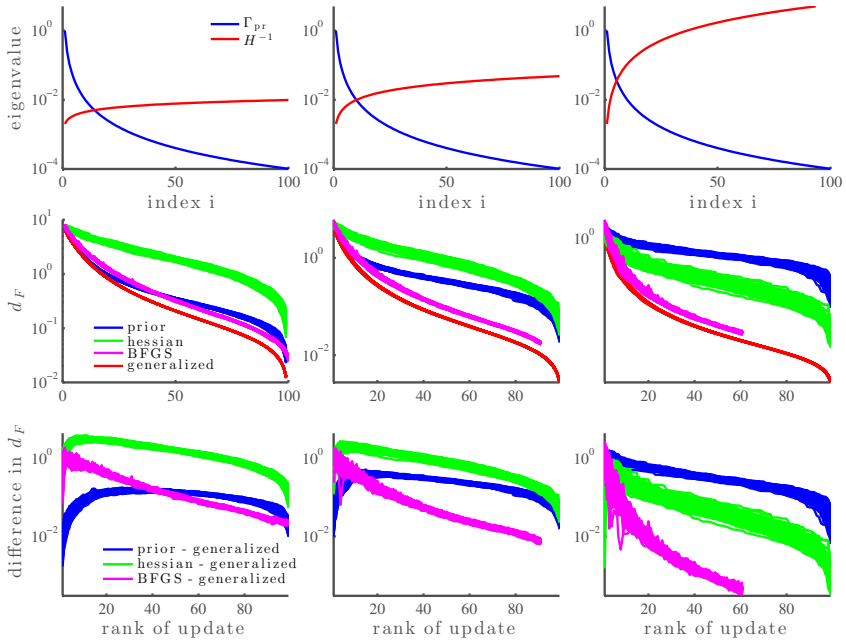

Figure 4.1: Top row: Eigenspectra of Γ_{pr} (blue) and H^{-1} (red) for three values for the decay rate of the eigenvalues of H: $\alpha = 0.345$ (left), $\alpha = 0.690$ (middle) and $\alpha = 1.724$ (right). Second row: Förstner distance between Γ_{pos} and its approximation versus the rank of the update for 100 realizations of Γ_{pr} and H using prior-based (blue), Hessian-based (green), BFGS-based (magenta) and optimal (red) updates. Bottom row: Differences of posterior covariance approximation error (measured with the Förstner metric) between the prior-based and optimal updates (blue), between the Hessian-based and optimal updates (green), and between the BFGS-based and optimal updates (magenta).

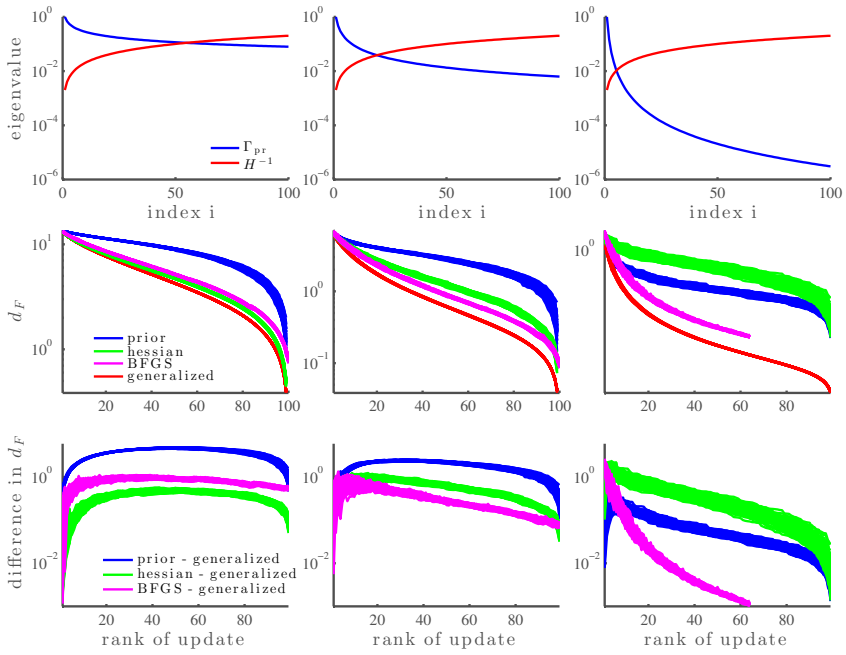

Figure 4.2: Analogous to Figure 4.1 but this time the spectrum of H is fixed, while that of Γ_{pr} has varying decay rates: $\widetilde{\alpha} = 0.552$ (left), $\widetilde{\alpha} = 1.103$ (middle) and $\widetilde{\alpha} = 2.759$ (right).

We model the absorption of an X-ray along a line ℓ_i using Beer's law:

$$I_d = I_s \exp\left(-\int_{\ell_i} x(s)ds\right), \tag{4.38}$$

where I_d and I_s are the intensities at the detector and at the source, respectively, and $x(s)$ is the density of the object at position s on the line ℓ_i. The computational domain is discretized into a grid and the density is assumed to be constant within each grid cell. The line integrals are approximated as

$$\int_{\ell_i} x(s)ds \approx \sum_{j=1}^{\#\text{ of cells}} g_{ij}x_j, \tag{4.39}$$

where g_{ij} is the length of the intersection between line ℓ_i and cell j, and x_j is the unknown density in cell j. The vector of absorptions along m lines can then be approximated as

$$I_d \approx I_s \exp\left(-Gx\right), \tag{4.40}$$

where I_d is the vector of m intensities at the detectors and $G = (g_{ij})$ is the $m \times n$ matrix of intersection lengths for each of the m lines. Even though the forward operator (4.40) is nonlinear, the inference problem can be recast in a linear fashion by taking logarithm of both sides of (4.40). This leads to the following linear model for the inversion: $y = Gx + \epsilon$, where the measurement vector is $y = -\log(I_d/I_s)$ and the measurement errors are assumed to be iid Gaussian, $\epsilon \sim \mathcal{N}(0, \sigma^2 I)$.

The setup for the inference problem, borrowed from [105], is as follows. The rectangular domain is discretized with an $n \times n$ grid. The true object

consists of three circular inclusions, each of uniform density, inside an annulus. Ten X-ray sources are positioned on one side of a circle, and each source sends a fan of 100 X-rays that are measured by detectors on the opposite side of the object. Here, the 10 sources are distributed evenly so that they form a total illumination angle of 90 degrees, resulting in a limited-angle CT problem. We use the exponential model (4.38) to generate synthetic data in a discretization-independent fashion by computing the exact intersections between the rays and the circular inclusions in the domain. Gaussian noise with standard deviation $\sigma = 0.002$ is added to the simulated data. The imaging setup and data from one source are illustrated in Figure 4.3.

The unknown density is estimated on a 128×128 grid. Thus the discretized vector x has length 16384, and direct computation of the posterior mean and the posterior covariance matrix, as well as generation of posterior samples, can be computationally nontrivial. To define the prior distribution, x is modeled as a discretized solution of a stochastic PDE of the form:

$$\gamma \left(\kappa^2 \mathcal{I} - \triangle\right) x(s) = \mathcal{W}(s), \qquad s \in \Omega, \tag{4.41}$$

where $\mathcal{W}$ is a white noise process, $\triangle$ is the Laplacian operator, and $\mathcal{I}$ is the identity operator. The solution of (4.41) is a Gaussian random field whose correlation length and variance are controlled by the free parameters κ and γ, respectively. A square root of the prior precision matrix of x (which is positive definite) can then be easily computed (see [137] for details). We use $\kappa = 10$ and $\gamma = \sqrt{800}$ in our simulations.

Our first task is to compute an optimal approximation of Γ_{pos} as a low-rank negative update of Γ_{pr} (cf. Theorem 1). Figure 4.4 (top row) shows the convergence of the approximate posterior variance as the rank of the update increases. The zero-rank update corresponds to Γ_{pr} (first column). For this formally 16384-dimensional problem, a good approximation of the posterior variance is achieved with a rank 200 update; hence the data are informative only on a low-dimensional subspace. The quality of the covariance matrix approximation is also reflected in the structure of samples drawn from the approximate posterior distributions (bottom row). All five of these samples are drawn using the same random seed and the exact posterior mean, so that all the differences observed are due to the approximation of Γ_{pos}. Already with a rank 100 update, the small-scale features of the approximate posterior sample match those of the exact posterior sample. In applications, agreement in this "eye-ball norm" is important. Of course, Theorem 1 also provides an exact formula for the error in the posterior covariance; this error is shown in the right panel of Figure 4.7 (blue curve).

Our second task is to assess the performances of the two optimal posterior mean approximations given in Section 4.4. We will use $\mu_{\text{pos}}^{(r)}(y)$ to denote the low-rank approximation and $\widehat{\mu}_{\text{pos}}^{(r)}(y)$ to denote the low-rank *update* approximation. Recall that both approximations are linear functions of the data y, given by $\mu_{\text{pos}}^{(r)}(y) = A^* y$ with $A^* \in \mathcal{A}_r$ and $\widehat{\mu}_{\text{pos}}^{(r)}(y) = \widehat{A}^* y$ with $\widehat{A}^* \in \widehat{\mathcal{A}}_r$, where the classes $\mathcal{A}_r$ and $\widehat{\mathcal{A}}_r$ are defined in (4.26). As in Section 4.4, we shall use $\mathcal{A}$ to denote either of the two classes.

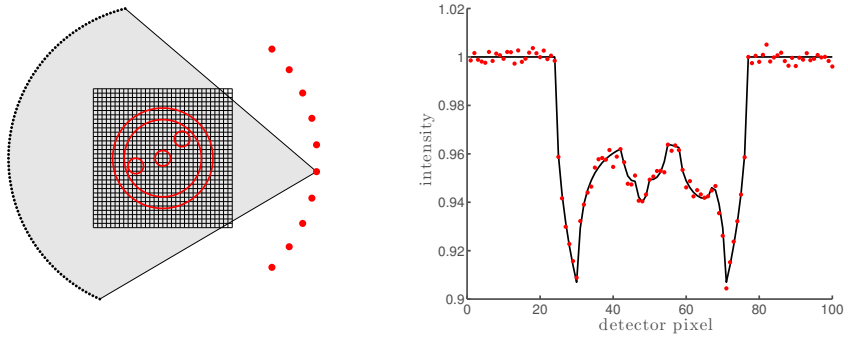

Figure 4.3: X-ray tomography problem. Left: Discretized domain, true object, sources (red dots), and detectors corresponding to one source (black dots). The fan transmitted by one source is illustrated in gray. The density of the object is 0.006 in the outer ring and 0.004 in the three inclusions; the background density is zero. Right: The true simulated intensity (black line) and noisy measurements (red dots) for one source.

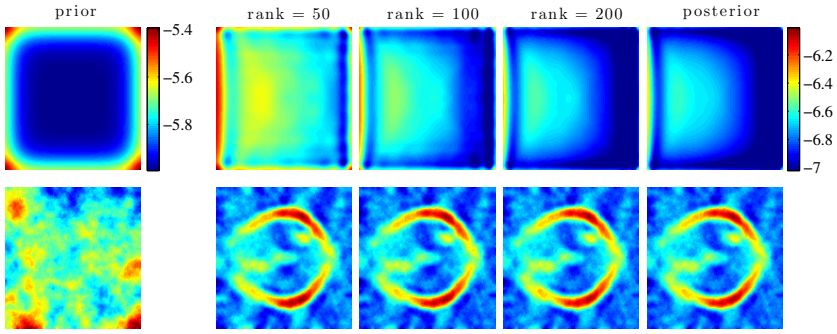

Figure 4.4: X-ray tomography problem. First column: Prior variance field, in log scale (top), and a sample drawn from the prior distribution (bottom). Second through last columns (left to right): Variance field, in log scale, of the approximate posterior as the rank of the update increases (top); samples from the corresponding approximate posterior distributions (bottom) assuming exact knowledge of the posterior mean.

Figure 4.5 shows the normalized error $\|\mu(y) - \mu_{\text{pos}}(y)\|_{\Gamma_{\text{pos}}^{-1}} / \|\mu_{\text{pos}}(y)\|_{\Gamma_{\text{pos}}^{-1}}$ for different approximations $\mu(y)$ of the true posterior mean $\mu_{\text{pos}}(y)$ and a fixed realization y of the data. The error is a function of the order r of the approximation class $\mathcal{A}$. Snapshots of $\mu(y)$ are shown along the two error curves. For reference, $\mu_{\text{pos}}(y)$ is also shown at the top. We see that the errors decrease monotonically, but that the low-rank approximation outperforms the low-rank update approximation for lower values of r. This is consistent with the discussion at the end of Section 4.4; the crossing point of the error curves is also consistent with that analysis. In particular, we expect the low-rank update approximation to outperform the low-rank approximation only when the approximation starts to include generalized eigenvalues of the pencil $(H, \Gamma_{\text{pr}}^{-1})$ that are less than one—i.e., once the approximations are no longer *under-resolved*. This can be confirmed by comparing Figure 4.5 with the decay of the generalized eigenvalues of the pencil $(H, \Gamma_{\text{pr}}^{-1})$ in the right panel of Figure 4.7 (blue curve).

On top of each snapshot in Figure 4.5, we show the *relative CPU time* required to compute the corresponding posterior mean approximation for each new realization of the data. The relative CPU time is defined as the time required to compute this approximation[7] divided by the time required to apply the posterior precision matrix to a vector. This latter operation is essential to computing the posterior mean via an iterative solver, such as a Krylov sub-

[7]This timing does not include the computation of (4.28) or (4.31), which should be regarded as *offline* steps. Here we report the time necessary to apply the optimal linear function to any new realization of the data, i.e., the *online* cost.

space method. These solvers are a standard choice for computing the posterior mean in large-scale inverse problems. Evaluating the ratio allows us to determine how many solver iterations could be performed with a computational cost roughly equal to that of approximating the posterior mean for a new realization of the data. Based on the reported times, a few observations can be made. First of all, as anticipated in Section 4.4, computing $\mu_{\text{pos}}^{(r)}(y)$ for any new realization of the data is faster than computing $\widehat{\mu}_{\text{pos}}^{(r)}(y)$. Second, obtaining an accurate posterior mean approximation requires roughly $r = 200$, and the relative CPU times for this order of approximation are 7.3 for $\mu_{\text{pos}}^{(r)}(y)$ and 29.0 for $\widehat{\mu}_{\text{pos}}^{(r)}(y)$; these are roughly the number of iterations of an iterative solver that one could take for equivalent computational cost. That is, the speedup of the posterior mean approximation compared to an iterative solver is not particularly dramatic in this case, because the forward model A is simply a sparse matrix that is cheap to apply. For the heat equation example discussed in Section 4.5.3, the situation is different.

Note that the above computational time estimates exclude other costs associated with iterative solvers. For instance, preconditioners are often applied; these significantly decrease the number of iterations needed for the solvers to converge but, on the other hand, increase the cost per iteration. A popular approach for solving the posterior mean efficiently is to use the prior covariance as the preconditioner [26]. In the limited-angle tomography problem, including the application of this preconditioner in the reference CPU time would reduce the relative CPU time of our $r = 200$ approximations to

0.48 for $\mu_{\text{pos}}^{(r)}(y)$ and 1.9 for $\widehat{\mu}_{\text{pos}}^{(r)}(y)$. That is, the cost of computing our approximations is roughly equal to *one iteration* of a prior-preconditioned iterative solver. The large difference compared to the case without preconditioning is due to the fact that applying the prior here is computationally much heavier than applying the forward model.

Figure 4.6 (left panel) shows unnormalized errors in the approximation of $\mu_{\text{pos}}(y)$,

$$\|e(y)\|^2_{\Gamma^{-1}_{\text{pos}}} = \|\mu^{(r)}_{\text{pos}}(y) - \mu_{\text{pos}}(y)\|^2_{\Gamma^{-1}_{\text{pos}}} \quad \text{and} \quad \|\widehat{e}(y)\|^2_{\Gamma^{-1}_{\text{pos}}} = \|\widehat{\mu}^{(r)}_{\text{pos}}(y) - \mu_{\text{pos}}(y)\|^2_{\Gamma^{-1}_{\text{pos}}}, \tag{4.42}$$

for the same realization of y used in Figure 4.5. In the same panel, we also show the expected values of these errors over the prior predictive distribution of y, which is exactly the r-dependent component of the Bayes risk given in Theorems 2 and 3. Both sets of errors decay with increasing r and show a similar crossover between the two approximation classes. But the particular error $\|e(y)\|^2_{\Gamma^{-1}_{\text{pos}}}$ departs consistently from its expectation; this is not unreasonable in general (the mean estimator has a nonzero variance), but the offset may be accentuated in this case because the data are generated from an image that is not drawn from the prior. (The right panel of Figure 4.6, which comes from Example 3, represents a contrasting case.)

By design, the posterior approximations described in this paper perform well when the data inform a low-dimensional subspace of the parameter space. To better understand this effect, we also consider a *full-angle* configuration of the tomography problem, wherein the sources and detectors are evenly spread

around the entire unknown object. In this case, the data are more informative than in the limited-angle configuration. This can be seen in the decay rate of the generalized eigenvalues of the pencil $(H, \Gamma_{\text{pr}}^{-1})$ in the center panel of Figure 4.7 (blue and red curves); eigenvalues for the full-angle configuration decay more slowly than for the limited-angle configuration. Thus, according to the optimal loss given in (4.10) (Theorem 1), the prior-to-posterior update in the full-angle case must be of greater rank than the update in the limited-angle case for any given approximation error. Also, good approximation of $\mu_{\text{pos}}(y)$ in the full-angle case requires higher order of the approximation class $\mathcal{A}$, as is shown in Figure 4.8. But because the data are strongly informative, they allow an almost perfect reconstruction of the underlying truth image. The relative CPU times are similar to the limited angle case: roughly 8 for $\mu_{\text{pos}}^{(r)}(y)$ and 14 for $\widehat{\mu}_{\text{pos}}^{(r)}(y)$. If preconditioning with the prior covariance is included in the reference CPU time calculation, the relative CPU times drop to 1.5 for $\mu_{\text{pos}}^{(r)}(y)$ and to 2.6 for $\widehat{\mu}_{\text{pos}}^{(r)}(y)$. We remark that in realistic applications of X-ray tomography, the limited angle setup is extremely common as it is cheaper and more flexible (yielding smaller and lighter devices) than a full-angle configuration.

4.5.3 Example 3: Heat equation

Our last example is the classic linear inverse problem of solving for the initial conditions of an inhomogeneous heat equation. Let $u(s,t)$ be the time dependent state of the heat equation on $s = (s_1, s_2) \in \Omega = [0,1]^2$, $t \geq 0$, and

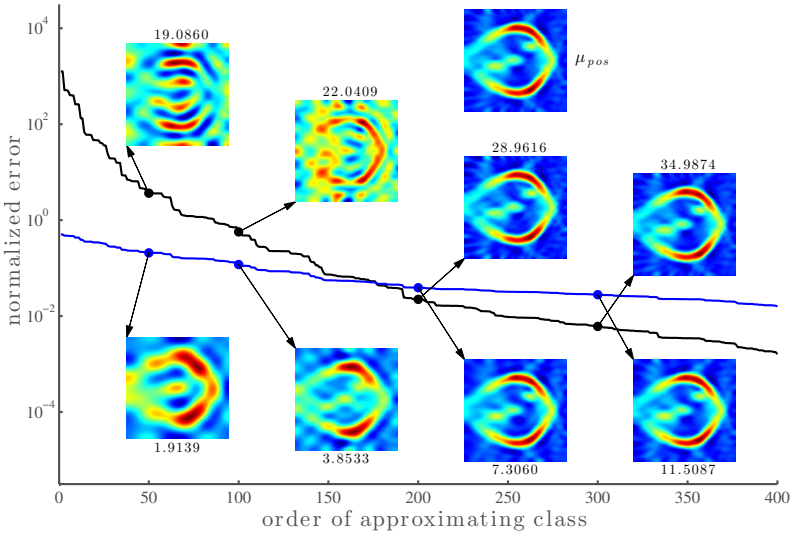

Figure 4.5: Limited-angle X-ray tomography: Comparison of the optimal posterior mean approximations, $\mu_{\text{pos}}^{(r)}(y)$ (blue) and $\widehat{\mu}_{\text{pos}}^{(r)}(y)$ (black) of $\mu_{\text{pos}}(y)$ for a fixed realization of the data y, as a function of the order r of the approximating classes $\mathcal{A}_r$ and $\widehat{\mathcal{A}}_r$, respectively. The normalized error for an approximation $\mu(y)$ is defined as $\|\mu(y) - \mu_{\text{pos}}(y)\|_{\Gamma_{\text{pos}}^{-1}} \,/\, \|\mu_{\text{pos}}(y)\|_{\Gamma_{\text{pos}}^{-1}}$. The numbers above or below the snapshots indicate the relative CPU time of the corresponding mean approximation—i.e., the time required to compute the approximation divided by the time required to apply the posterior precision matrix to a vector.

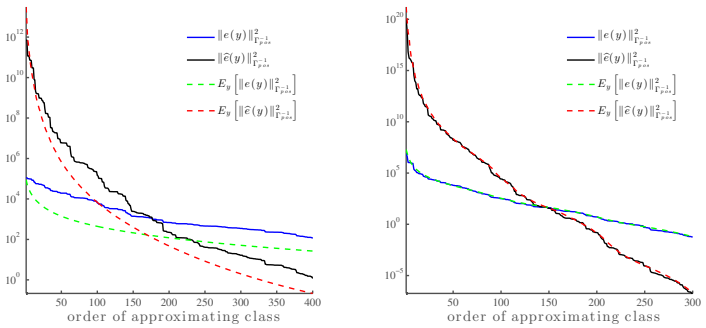

Figure 4.6: The errors $\|e(y)\|^2_{\Gamma^{-1}_{\text{pos}}}$ (blue) and $\|\widehat{e}(y)\|^2_{\Gamma^{-1}_{\text{pos}}}$ (black) defined by (4.42), and their expected values in green and red, respectively; for Example 2 (left panel) and Example 3 (right panel).

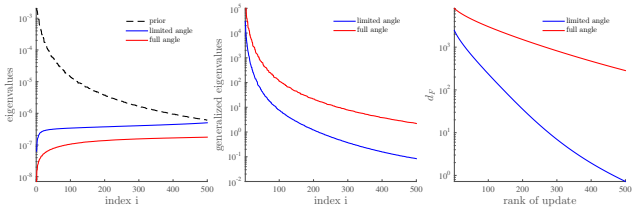

Figure 4.7: Left: Leading eigenvalues of Γ_{pr} and H^{-1} in the limited-angle and full-angle X-ray tomography problems. Center: Leading generalized eigenvalues of the pencil $(H, \Gamma^{-1}_{\text{pr}})$ in the limited-angle (blue) and full-angle (red) cases. Right: $d_{\mathcal{F}}(\Gamma_{\text{pos}}, \widehat{\Gamma}_{\text{pos}})$ as a function of the rank of the update $KK^\top$, with $\widehat{\Gamma}_{\text{pos}} = \Gamma_{\text{pr}} - KK^\top$, in the limited-angle (blue) and full-angle (red) cases.

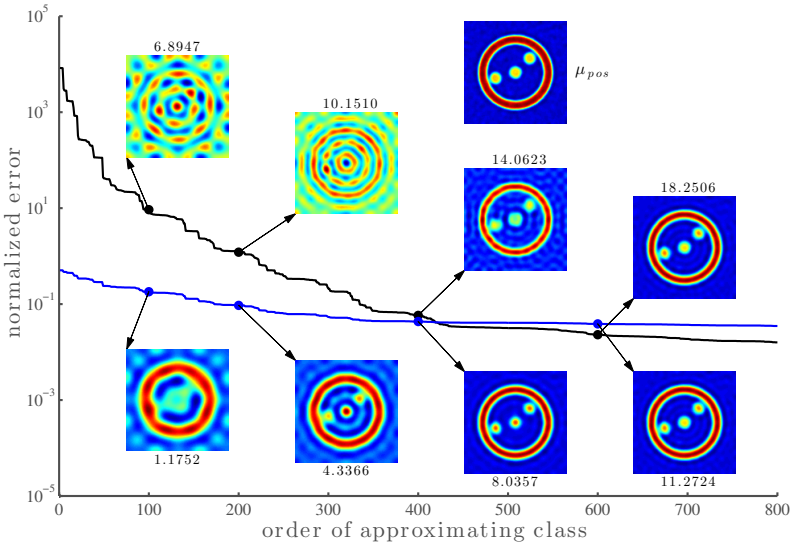

Figure 4.8: Same as Figure 4.5, but for full-angle X-ray tomography (sources and receivers spread uniformly around the entire object).

let $\kappa(s)$ be the heat conductivity field. Given initial conditions $u_0(s) = u(s, 0)$, the state evolves in time according to the linear heat equation:

$$\begin{aligned} \frac{\partial u(s,t)}{\partial t} &= -\nabla \cdot (\kappa(s) \nabla u(s,t)), \qquad s \in \Omega,\ t > 0, \\ \kappa(s) \nabla u(s,t) \cdot n(s) &= 0, \qquad s \in \partial\Omega,\ t > 0, \end{aligned} \tag{4.43}$$

where $n(s)$ denotes the outward-pointing unit normal at $s \in \partial\Omega$. We place $n_s = 81$ sensors at the locations $s_1, \ldots, s_{n_s}$, uniformly spaced within the lower left quadrant of the spatial domain, as illustrated by the black dots in Figure 4.9. We use a finite dimensional discretization of the parameter space based on the finite element method on a regular 100×100 grid, $\{s_i'\}$. Our goal is to infer the vector $x = (u_0(s_i'))$ of initial conditions on the grid. Thus, the

dimension of the parameter space for the inference problem is $n = 10^4$. We use data measured at 50 discrete times $t = t_1, t_2, \ldots, t_{50}$, where $t_i = i\triangle t$, and $\triangle t = 2 \times 10^{-4}$. At each time t_i, pointwise observations of the state u are taken at these sensors, i.e.,

$$d_i = \mathcal{C}u(s, t_i), \tag{4.44}$$

where $\mathcal{C}$ is the observation operator that maps the function $u(s, t_i)$ to $d = (u(s_1, t_i), \ldots, u(s_n, t_i))^\top$. The vector of observations is then $d = [d_1; d_2; \ldots; d_{50}]$. The noisy data vector is $y = d + \varepsilon$, where $\varepsilon \sim \mathcal{N}(0, \sigma^2 I)$ and $\sigma = 10^{-2}$. Note that the data are a linear function of the initial conditions, perturbed by Gaussian noise. Thus the data can be written as:

$$y = Gx + \varepsilon, \qquad \varepsilon \sim \mathcal{N}(0, \sigma^2 I). \tag{4.45}$$

where G is a linear map defined by the composition of the forward model (4.43) with the observation operator (4.44), both linear.

We generate synthetic data by evolving the initial conditions shown in Figure 4.9. This "true" value of the inversion parameters x is a discretized realization of a Gaussian process satisfying an SPDE of the same form used in the previous tomography example, but now with a non-stationary permeability field. In other words, the truth is a draw from the prior in this example (unlike in the previous example), and the prior Gaussian process satisfies the following SPDE:

$$\gamma \left(\kappa^2 \mathcal{I} - \nabla \cdot \mathbf{c}(s) \nabla \right) x(s) = \mathcal{W}(s) \qquad s \in \Omega, \tag{4.46}$$

where $\mathbf{c}(s)$ is the space-dependent permeability tensor.

Figure 4.10 and the right panel in Figure 4.6 show our numerical results. They have the same interpretations as Figures 4.5 and 4.6 in the tomography example. The trends in the figures are consistent with those encountered in the previous example and confirm the good performance of the optimal low-rank approximation. Notice that in Figures 4.10 and 4.6 the approximation of the posterior mean appears to be nearly perfect (visually) once the error curves for the two approximations cross. This is somewhat expected from the theory since we know that the crossing point should occur when the approximations start to use eigenvalues of the pencil $(H, \Gamma_{\text{pr}}^{-1})$ that are less than one—that is, once we have exhausted directions in the parameter space where the data are more constraining than the prior.

Again, we report the relative CPU time for each posterior mean approximation above/below the corresponding snapshot in Figure 4.10. The results differ significantly from the tomography example. For instance, at order $r = 200$, which yields approximations that are visually indistinguishable from the true mean, the relative CPU times are 0.001 for $\mu_{\text{pos}}^{(r)}(y)$ and 0.53 for $\widehat{\mu}_{\text{pos}}^{(r)}(y)$. Therefore we can compute an accurate mean approximation for a new realization of the data much more quickly than taking one iteration of an iterative solver. Recall that, consistent with the setting described at the start of Section 4.4, this is a comparison of *online* times, after the matrices (4.28) or (4.31) have been precomputed. The difference between this case and tomography example of Section 4.5.2 is due to the higher CPU cost of applying the forward and adjoint models for the heat equation—solving a time

dependent PDE versus applying a sparse matrix. Also, because the cost of applying the prior covariance is negligible compared to that of the forward and adjoint solves in this example, preconditioning the iterative solver with the prior would not strongly affect the reported relative CPU times, unlike the tomography example.

Figure 4.11 illustrates some important directions characterizing the heat equation inverse problem. The first two columns show the four leading eigenvectors of, respectively, Γ_{pr} and H. Notice that the support of the eigenvectors of H concentrates around the sensors. The third column shows the four leading directions ($\widehat{w}_i$) defined in Theorem 1. These directions define the optimal prior-to-posterior covariance matrix update (cf. (4.9)). This update of Γ_{pr} is *necessary* to capture directions ($\widetilde{w}_i$) of greatest relative difference between prior and posterior variance (cf. Corollary 1). The four leading directions ($\widetilde{w}_i$) are shown in the fourth column. The support of these modes is again concentrated around the sensors, which intuitively makes sense as these are directions of greatest variance reduction.

4.6 Conclusions

This paper has presented and characterized optimal approximations of the Bayesian solution of linear inverse problems, with Gaussian prior and noise distributions defined on finite-dimensional spaces. In a typical large-scale inverse problem, observations may be informative—relative to the prior—only on a low-dimensional subspace of the parameter space. Our approximations

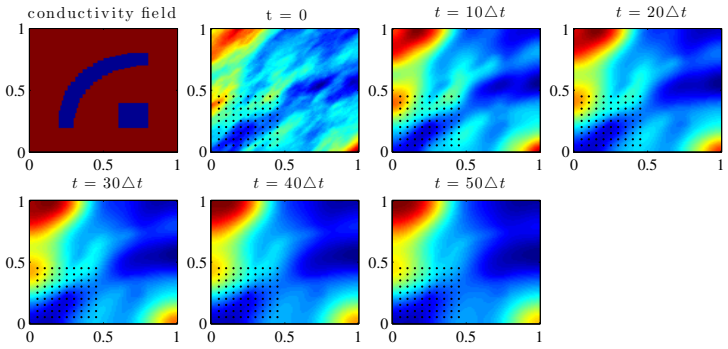

Figure 4.9: Heat equation (Example 3). Initial condition (top left) and several snapshots of the states at different times. Black dots indicate sensor locations.

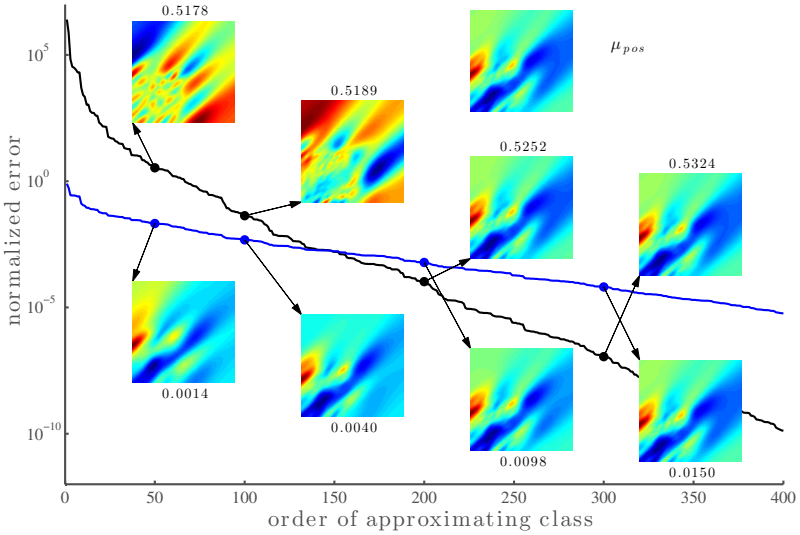

Figure 4.10: Same as Figure 4.5, but for Example 3 (initial condition inversion for the heat equation).

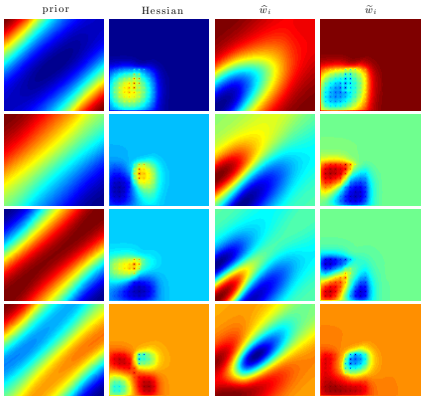

Figure 4.11: Heat equation (Example 3). First column: Four leading eigenvectors of Γ_{pr}. Second column: Four leading eigenvectors of H. Third column: Four leading directions ($\widehat{w}_i$) (cf. (4.9)). Fourth column: Four leading directions ($\widetilde{w}_i$) (cf. Corollary 1)

therefore identify and exploit low-dimensional structure in the *update* from prior to posterior.

We have developed two types of optimality results. In the first, the posterior covariance matrix is approximated as a low-rank negative semidefinite update of the prior covariance matrix. We describe an update of this form that is optimal with respect to a broad class of loss functions between covariance matrices, exemplified by the Förstner metric [75] for symmetric positive definite matrices. We argue that this is the appropriate class of loss functions with which to evaluate approximations of the posterior covariance matrix, and show that optimality in such metrics identifies directions in parameter space along which the posterior variance is reduced the most, relative to the prior. Optimal low-rank updates are derived from a generalized eigendecomposition

of the pencil defined by the minus log-likelihood Hessian and the prior precision matrix. These updates have been proposed in previous work [73], but our work complements these efforts by characterizing the optimality of the resulting approximations. Under the assumption of exact knowledge of the posterior mean, our results extend to optimality statements between the associated distributions (e.g., optimality in the Hellinger distance and in the Kullback-Leibler divergence). Second, we have developed fast approximations of the posterior mean that are useful when repeated evaluations thereof are required for multiple realizations of the data (e.g., in an online inference setting). These approximations are optimal in the sense that they minimize the Bayes risk for squared-error loss induced by the posterior precision matrix. The most computationally efficient of these approximations expresses the posterior mean as the product of a single low-rank matrix with the data. We have demonstrated the covariance and mean approximations numerically on a variety of inverse problems: synthetic problems constructed from random Hessian and prior covariance matrices; an X-ray tomography problem with different observation scenarios; and inversion for the initial condition of a heat equation, with localized observations and a non-stationary prior.

This work has several possible extensions of interest, some of which are already part of ongoing research. First, it is natural to generalize the present approach to infinite-dimensional parameter spaces endowed with Gaussian priors. This setting is essential to understanding and formalizing Bayesian inference over function spaces [33, 189]. Here, by analogy with the current results,

one would expect the posterior covariance operator to be well approximated by a finite-rank negative perturbation of the prior covariance operator. A further extension could allow the data to become infinite-dimensional as well. Another important task is to generalize the present methodology to inverse problems with nonlinear forward models. One approach for doing so is presented in [57]; other approaches are certainly possible. Yet another interesting research topic is the study of analogous approximation techniques for sequential inference. We note that the assimilation step in a linear (or linearized) data assimilation scheme can be already tackled within the framework presented here. But the nonstationary setting, where inference is interleaved with evolution of the state, introduces the possibility for even more tailored and structure-exploiting approximations.

4.7 Technical results

Here we collect the proofs and other technical results necessary to support the statements made in the previous sections.

We start with an auxiliary approximation result that plays an important role in our analysis. Given a semi-positive definite diagonal matrix D, we seek an approximation of $D + I$ by a rank r perturbation of the identity, $UU^\top + I$, that minimizes a loss function from the class $\mathcal{L}$ defined in (4.6). The following lemma shows that the optimal solution $\widehat{U}\widehat{U}^\top$ is simply the best rank r approximation of the matrix D in the Frobenius norm.

Lemma 2 (Approximation lemma)**.** *Let* $D = \operatorname{diag}\{d_1^2, \ldots, d_n^2\}$*, with* $d_i^2 \geq d_{i+1}^2$*, and* $L \in \mathcal{L}$*. Define the functional* $\mathcal{J} : \mathbb{R}^{n\times r} \to \mathbb{R}$*, as:* $\mathcal{J}(U) = L(UU^\top + I, D + I) = \sum_i f(\sigma_i)$*, where* (σ_i) *are the generalized eigenvalues of the pencil* $(UU^\top + I, D + I)$ *and* $f \in \mathcal{U}$*. Then:*

(i) There is a minimizer, $\widehat{U}$*, of* $\mathcal{J}$ *such that*

$$\widehat{U}\widehat{U}^\top = \sum_{i=1}^{r} d_i^2\, e_i e_i^\top. \tag{4.47}$$

where (e_i) *are the columns of the identity matrix.*

(ii) If the first r *eigenvalues of* D *are distinct, then any minimizer of* $\mathcal{J}$ *satisfies* (4.47)*.*

Proof. The idea is to apply [131, Theorem 1.1] to the functional $\mathcal{J}$. To this end, we notice that $\mathcal{J}$ can be equivalently written as: $\mathcal{J}(U) = F \circ \rho_n \circ g(U)$, where: $F : \mathbb{R}_+^n \to \mathbb{R}$ is of the form $F(x) = \sum_{i=1}^n f(x_i)$; ρ_n denotes a function that maps an $n \times n$ SPD matrix A to its eigenvalues $\sigma = (\sigma_i)$ (i.e., $\rho_n(A) = \sigma$ and since F is a symmetric function, the order of the eigenvalues is irrelevant); and the mapping g is given by: $g(U) = (D+I)^{-1/2}(UU^\top+I)(D+I)^{-1/2}$, for all $U \in \mathbb{R}^{n\times r}$. Since the function $F \circ \rho_n$ satisfies the hypotheses in [131, Theorem 1.1], $F\circ\rho_n$ is differentiable at the SPD matrix X if and only if F is differentiable at $\rho_n(X)$, in which case $(F \circ \rho_n)'(X) = ZS_\sigma Z^\top$, where

$$S_\sigma = \operatorname{diag}[\, F'(\rho_n(X))\,] = \operatorname{diag}\{f'(\sigma_1), \ldots, f'(\sigma_n)\},$$

and Z is an orthogonal matrix such that $X = Z\,\mathrm{diag}[\,\rho_n(X)\,]Z^\top$. Using the chain rule, we obtain

$$\frac{\partial \mathcal{J}(U)}{\partial U_{ij}} = \mathrm{tr}\left(Z S_\sigma Z^\top \frac{\partial g(U)}{\partial U_{ij}}\right),$$

which leads to the following gradient of $\mathcal{J}$ at U:

$$\mathcal{J}'(U) = 2(D+I)^{-1/2} Z S_\sigma (D+I)^{-1/2} Z^\top U = 2\,W S_\sigma W^\top U,$$

where the orthogonal matrix Z is such that the matrix $W = (D+I)^{-1/2}Z$ satisfies

$$(UU^\top + I)W = (D+I)W\Upsilon_\sigma \tag{4.48}$$

with $\Upsilon_\sigma = \mathrm{diag}(\sigma)$. Now we show that the functional $\mathcal{J}$ is coercive. Let (U_k) be a sequence of matrices such that $\|U_k\|_F \to \infty$. Hence, $\sigma_{max}(g(U_k)) \to \infty$ and so does $\mathcal{J}$ since:

$$\mathcal{J}(U_k) \geq f(\sigma_{max}(g(U_k))) + (n-1)f(1)$$

and $f(x) \to \infty$ as $x \to \infty$. Thus, $\mathcal{J}$ is a differentiable coercive functional, and has a global minimizer $\widehat{U}$ with zero gradient:

$$\mathcal{J}'(\widehat{U}) = 2W S_\sigma W^\top \widehat{U} = 0. \tag{4.49}$$

However, since $f \in \mathcal{U}$, $f'(x) = 0$ iff $x = 1$. It follows that condition (4.49) is equivalent to

$$(I - \Upsilon_\sigma)W^\top \widehat{U} = 0. \tag{4.50}$$

(4.48) and (4.50) give $\widehat{U}\widehat{U}^\top - D = W^{-\top}\Upsilon^{-1}(\Upsilon - I)W^\top$, and right-multiplication by $\widehat{U}\widehat{U}^\top$ then yields:

$$D\,(\widehat{U}\widehat{U}^\top) = (\,\widehat{U}\widehat{U}^\top)^2. \tag{4.51}$$

In particular, if u is an eigenvector of $\widehat{U}\widehat{U}^\top$ with nonzero eigenvalue α, then u is an eigenvector of D, $Du = \alpha u$, and thus $\alpha = d_i^2 > 0$ for some i. Thus, any solution of (4.51) is such that:

$$\widehat{U}\widehat{U}^\top = \sum_{i=1}^{r_k} d_{k_i}^2 e_{k_i} e_{k_i}^\top, \tag{4.52}$$

for some subsequence (k_ℓ) of $\{1, \ldots, n\}$ and rank $r_k \leq r$. Notice that any $\widehat{U}$ satisfying (4.51) is also a critical point according to (4.50). From (4.52) we also find that $g(\widehat{U})$ is a diagonal matrix,

$$g(\widehat{U}) = (D + I)^{-1} \left(\sum_{i=1}^{r_k} d_{k_i}^2 e_{k_i} e_{k_i}^\top + I \right).$$

The diagonal entries σ_i, which are the eigenvalues of $g(\widehat{U})$, are given by $\sigma_i = 1$ if $i = k_\ell$ for some $\ell \leq r_k$, or $\sigma_i = 1/(1 + d_i^2)$ otherwise. In either case, we have $0 < \sigma_i \leq 1$ and the monotonicity of f implies that $\mathcal{J}(\widehat{U})$ is minimized by the subsequence $k_1 = 1, \ldots, k_r = r$, and by the choice $r_k = r$. This proves (4.47). It is clear that if the first r eigenvalues of D are distinct, then any minimizer of $\mathcal{J}$ satisfies (4.47). □

Most of the objective functions we consider have the same structure as the loss function $\mathcal{J}$. Hence, the importance of Lemma 2.

The next lemma shows that searching for a negative update of Γ_{pr} is equivalent to looking for a positive update of the prior precision matrix. In

particular, the lemma provides a bijection between the two approximation classes, $\mathcal{M}_r$ and $\mathcal{M}_r^{-1}$, defined by (4.4) and (4.12). In what follows, S_{pr} is any square root of the prior covariance matrix such that $\Gamma_{\text{pr}} = S_{\text{pr}} S_{\text{pr}}^\top$.

Lemma 3 (Prior updates)**.** *For any negative semidefinite update of* Γ_{pr}*,* $\widehat{\Gamma}_{\text{pos}} = \Gamma_{\text{pr}} - KK^\top$ *with* $\widehat{\Gamma}_{\text{pos}} \succ 0$*, there is a matrix* U *(of the same rank as* K*) such that* $\widehat{\Gamma}_{\text{pos}} = \left(\Gamma_{\text{pr}}^{-1} + UU^\top\right)^{-1}$*. The converse is also true.*

Proof. Let $ZDZ^\top = S_{\text{pr}}^{-1} K K^\top S_{\text{pr}}^{-\top}$, $D = \text{diag}\{d_i^2\}$, be a reduced SVD of $S_{\text{pr}}^{-1} K K^\top S_{\text{pr}}^{-\top}$. Since $\widehat{\Gamma}_{\text{pos}} \succ 0$ by assumption, we must have $d_i^2 < 1$ for all i, and we may thus define $U = S_{\text{pr}}^{-\top} Z D^{1/2} (I - D)^{-1/2}$. By Woodbury's identity:

$$\left(\Gamma_{\text{pr}}^{-1} + UU^\top\right)^{-1} = \Gamma_{\text{pr}} - \Gamma_{\text{pr}} U \left(I + U^\top \Gamma_{\text{pr}}^{-1} U\right)^{-1} U^\top \Gamma_{\text{pr}} = \Gamma_{\text{pr}} - KK^\top = \widehat{\Gamma}_{\text{pos}}.$$

Conversely, given a matrix U, we use again Woodbury's identity to write $\widehat{\Gamma}_{\text{pos}}$ as a negative semidefinite update of Γ_{pr}: $\widehat{\Gamma}_{\text{pos}} = \Gamma_{\text{pr}} - KK^\top \succ 0$. □

Now we prove our main result on approximations of the posterior covariance matrix.

Proof of Theorem 1. Given a loss function $L \in \mathcal{L}$, our goal is to minimize:

$$L(\Gamma_{\text{pos}}, \widehat{\Gamma}_{\text{pos}}) = \sum_i f(\sigma_i) \tag{4.53}$$

over $K \in \mathbb{R}^{n \times r}$ subject to the constraint $\widehat{\Gamma}_{\text{pos}} = \Gamma_{\text{pr}} - KK^\top \succ 0$, where (σ_i) are the generalized eigenvalues of the pencil $(\Gamma_{\text{pos}}, \widehat{\Gamma}_{\text{pos}})$ and f belongs to the class $\mathcal{U}$ defined by Eq. (4.7). We also write $\sigma_i(\Gamma_{\text{pos}}, \widehat{\Gamma}_{\text{pos}})$ to specify the pencil

corresponding to the eigenvalues. By Lemma 3, the optimization problem is equivalent to finding a matrix, $U \in \mathbb{R}^{n\times r}$, that minimizes (4.53) subject to $\widehat{\Gamma}_{\text{pos}}^{-1} = \Gamma_{\text{pr}}^{-1} + UU^\top$. Observe that (σ_i) are also the eigenvalues of the pencil $(\widehat{\Gamma}_{\text{pos}}^{-1}, \Gamma_{\text{pos}}^{-1})$.

Let $WDW^\top = S_{\text{pr}}^\top H\, S_{\text{pr}}$ with $D = \text{diag}\{\delta_i^2\}$, be an SVD of $S_{\text{pr}}^\top H\, S_{\text{pr}}$. Then, by the invariance properties of the generalized eigenvalues we have:

$$\sigma_i(\widehat{\Gamma}_{\text{pos}}^{-1},\, \Gamma_{\text{pos}}^{-1}) \;\; = \;\; \sigma_i(\, W^\top S_{\text{pr}}^\top\, \widehat{\Gamma}_{\text{pos}}^{-1}\, S_{\text{pr}} W\,,\, W^\top S_{\text{pr}}^\top\, \Gamma_{\text{pos}}^{-1}\, S_{\text{pr}} W\,) = \sigma_i(\, ZZ^\top + I,\, D + I\,),$$

where $Z = W^\top S_{\text{pr}}^\top U$. Therefore, our goal reduces to finding a matrix, $Z \in \mathbb{R}^{n\times r}$, that minimizes (4.53) with (σ_i) being the generalized eigenvalues of the pencil $(\, ZZ^\top + I,\, D + I\,)$. Applying Lemma 2 leads to the simple solution: $ZZ^\top = \sum_{i=1}^r \delta_i^2\, e_i e_i^\top$, where (e_i) are the columns of the identity matrix. In particular, the solution is unique if the first r eigenvalues of $S_{\text{pr}}^\top H\, S_{\text{pr}}$ are distinct. The corresponding approximation $UU^\top$ is then

$$UU^\top = S_{\text{pr}}^{-\top}\, WZZ^\top W^\top S_{\text{pr}}^{-1} = \sum_{i=1}^{r} \delta_i^2\, \widetilde{w}_i \widetilde{w}_i^\top, \tag{4.54}$$

where $\widetilde{w}_i = S_{\text{pr}}^{-\top} w_i$ and w_i is the ith column of W. Woodbury's identity gives the corresponding negative update of Γ_{pr} as:

$$\widehat{\Gamma}_{\text{pos}} = \Gamma_{\text{pr}} - KK^\top, \quad KK^\top = \sum_{i=1}^{r} \delta_i^2 \left(1 + \delta_i^2\right)^{-1} \widehat{w}_i \widehat{w}_i^\top \tag{4.55}$$

with $\widehat{w}_i = S_{\text{pr}} w_i$. Now, it suffices to note that the couples $(\delta_i^2, \widehat{w}_i)$ defined here are precisely the generalized eigenpairs of the pencil $(H, \Gamma_{\text{pr}}^{-1})$. At optimality, $\sigma_i = 1$ for $i \le r$ and $\sigma_i = (1 + \delta_i^2)^{-1}$ for $i > r$, proving (4.10). $\square$

Before proving Lemma 1, we recall that the Kullback-Leibler (K-L) divergence and the Hellinger distance between two multivariate Gaussians, $\nu_1 = \mathcal{N}(\mu, \Sigma_1)$ and $\nu_2 = \mathcal{N}(\mu, \Sigma_2)$, with the same mean and full rank covariance matrices are given, respectively, by [161]:

$$D_{\mathrm{KL}}\left(\nu_1 \| \, \nu_2\right) = \frac{1}{2}\left[\mathrm{trace}\left(\Sigma_2^{-1}\Sigma_1\right) - \mathrm{rank}(\Sigma_1) - \ln\left(\frac{\det(\Sigma_1)}{\det(\Sigma_2)}\right)\right] \tag{4.56}$$

$$d_{\mathrm{Hell}}\left(\nu_1, \nu_2\right) = \sqrt{1 - \frac{|\Sigma_1|^{1/4}\,|\Sigma_2|^{1/4}}{|\frac{1}{2}\Sigma_1 + \frac{1}{2}\Sigma_2|^{1/2}}}. \tag{4.57}$$

Proof of Lemma 1. By (4.56), the K-L divergence between the posterior $\nu_{\mathrm{pos}}(y)$ and the Gaussian approximation $\widehat{\nu}_{\mathrm{pos}}(y)$ can be written in terms of the generalized eigenvalues of the pencil $(\Gamma_{\mathrm{pos}}, \widehat{\Gamma}_{\mathrm{pos}})$ as:

$$D_{\mathrm{KL}}\left(\nu_{\mathrm{pos}}(y) \| \, \widehat{\nu}_{\mathrm{pos}}(y)\right) = \sum_i \left(\, \sigma_i - \ln \sigma_i - 1 \,\right)/2,$$

and since $f(x) = \left(x - \ln x - 1\right)/2$ belongs to $\mathcal{U}$, we see that the K-L divergence is a loss function in the class $\mathcal{L}$ defined by (4.6). Hence, Theorem 1 applies and the equivalence between the two approximations follows trivially. An analogous argument holds for the Hellinger distance. The squared Hellinger distance between $\nu_{\mathrm{pos}}(y)$ and $\widehat{\nu}_{\mathrm{pos}}(y)$ can be written in terms of the generalized eigenvalues, (σ_i), of the pencil $(\Gamma_{\mathrm{pos}}, \widehat{\Gamma}_{\mathrm{pos}})$, as:

$$d_{\mathrm{Hell}}\left(\nu_{\mathrm{pos}}(y), \widehat{\nu}_{\mathrm{pos}}(y)\right)^2 = 1 - 2^{1/2}\prod_i \sigma_i^{1/4}\left(1 + \sigma_i\right)^{-1/2}. \tag{4.58}$$

Minimizing (4.58) is equivalent to maximizing $\prod_i \sigma_i^{1/4}\left(1 + \sigma_i\right)^{-1/2}$, which in turn is equivalent to minimizing the functional:

$$L(\Gamma_{\mathrm{pos}}, \widehat{\Gamma}_{\mathrm{pos}}) = -\sum_i \ln(\, \sigma_i^{1/4}(1 + \sigma_i)^{-1/2}\,) = \sum_i \ln(\, 2 + \sigma_i + 1/\sigma_i\,)/4 \tag{4.59}$$

Since $f(x) = \ln(\,2 + x + 1/x\,)/4$ belongs to $\mathcal{U}$, Theorem 1 can be applied once again. $\square$

Proof of Corollary 1. The proofs of parts (i) and (ii) were already given in the proof of Theorem 1. Part (iii) holds because,

$$\begin{aligned}(1+\delta_i^2)\Gamma_{\rm pos}\,\widetilde{w}_i &= (1+\delta_i^2)(H+\Gamma_{\rm pr}^{-1})^{-1}S_{\rm pr}^{-\top}w_i \\ &= (1+\delta_i^2)\,S_{\rm pr}(S_{\rm pr}^{\top}H\,S_{\rm pr}+I)^{-1}w_i = S_{\rm pr}\,w_i = \Gamma_{\rm pr}\,\widetilde{w}_i,\end{aligned}$$

because w_i is an eigenvector of $(\,S_{\rm pr}^{\top}H\,S_{\rm pr}+I\,)^{-1}$ with eigenvalue $(1+\delta_i^2)^{-1}$ as shown in the proof of Theorem 1. $\square$

Now we turn to optimality results for approximations of the posterior mean. In what follows, let $S_{\rm pr}$, $S_{\rm obs}$, $S_{\rm pos}$, and $S_{\rm y}$ be the matrix square roots of, respectively, $\Gamma_{\rm pr}$, $\Gamma_{\rm obs}$, $\Gamma_{\rm pos}$, and $\Gamma_{\rm y} := \Gamma_{\rm obs} + G\,\Gamma_{\rm pr}\,G^{\top}$ such that $\Gamma = S\,S^{\top}$ (i.e., possibly non-symmetric square roots).

Equation (4.25) shows that, to minimize $\mathbb{E}(\,\|Ay-x\|^2_{\Gamma_{\rm pos}^{-1}}\,)$ over $A \in \mathcal{A}$, we need only to minimize $\mathbb{E}(\,\|\,Ay-\mu_{\rm pos}(y)\,\|^2_{\Gamma_{\rm pos}^{-1}}\,)$. Furthermore, since $\mu_{\rm pos}(y) = \Gamma_{\rm pos}\,G^{\top}\Gamma_{\rm obs}^{-1}\,y$, it follows that

$$\mathbb{E}(\,\|\,Ay-\mu_{\rm pos}(y)\,\|^2_{\Gamma_{\rm pos}^{-1}}\,) = \|\,S_{\rm pos}^{-1}\,(A-\Gamma_{\rm pos}\,G^{\top}\Gamma_{\rm obs}^{-1})\,S_{\rm y}\,\|_F^2, \tag{4.60}$$

We are therefore led to the following optimization problem:

$$\min_{A\in\mathcal{A}} \|\,S_{\rm pos}^{-1}A\,S_{\rm y} - S_{\rm pos}^{\top}G^{\top}\Gamma_{\rm obs}^{-1}\,S_{\rm y}\,\|_F. \tag{4.61}$$

The following result shows that an SVD of the matrix $S_{\widehat{H}} := S_{\rm pr}^{\top}\,G^{\top}S_{\rm obs}^{-\top}$ can be used to obtain simple expressions for the square roots of $\Gamma_{\rm pos}$ and $\Gamma_{\rm y}$.

Lemma 4 (Square roots)**.** *Let $WDV^\top$ be an SVD of $S_{\widehat{H}} = S_{\rm pr}^\top G^\top S_{\rm obs}^{-\top}$. Then:*

$$S_{\rm pos} = S_{\rm pr}\, W(\, I + DD^\top\,)^{-1/2}\, W^\top \tag{4.62}$$

$$S_{\rm y} = S_{\rm obs}\, V(\, I + D^\top D\,)^{1/2}\, V^\top \tag{4.63}$$

are square roots of $\Gamma_{\rm pos}$ and $\Gamma_{\rm y}$.

Proof. We can rewrite $\Gamma_{\rm pos} = (\, G^\top \Gamma_{\rm obs}^{-1} G + \Gamma_{\rm pr}^{-1}\,)^{-1}$ as

$$\begin{aligned}\Gamma_{\rm pos} &= S_{\rm pr}\,(\, S_{\widehat{H}}\, S_{\widehat{H}}^\top + I\,)^{-1}\, S_{\rm pr}^\top = S_{\rm pr}\, W(\, DD^\top + I\,)^{-1} W^\top S_{\rm pr}^\top \\ &= [\, S_{\rm pr}\, W(\, DD^\top + I\,)^{-1/2}\, W^\top\,]\,[\, S_{\rm pr}\, W(\, DD^\top + I\,)^{-1/2}\, W^\top\,]^\top,\end{aligned}$$

which proves (4.62). The proof of (4.63) follows similarly using: $S_{\widehat{H}}^\top S_{\widehat{H}} = S_{\rm obs}^{-1} G\, \Gamma_{\rm pr}\, G^\top \Gamma_{\rm obs}^{-\top}$. □

In the next two proofs we use $(C)_r$ to denote a rank r approximation of the matrix C in the Frobenius norm.

Proof of Theorem 2. By [77, Theorem 2.1], an optimal $A \in \mathcal{A}_r$ is given by:

$$A = S_{\rm pos}\, \left(\, S_{\rm pos}^\top G^\top \Gamma_{\rm obs}^{-1}\, S_{\rm y}\,\right)_r\, S_{\rm y}^{-1}. \tag{4.64}$$

Now, we need some computations to show that (4.64) is equivalent to (4.28). Using (4.62) and (4.63) we find $S_{\rm pos}^\top G^\top \Gamma_{\rm obs}^{-1}\, S_{\rm y} = W(I + DD^\top)^{-1/2} D\,(I + D^\top D)^{1/2}\, V^\top$, and therefore $(\, S_{\rm pos}^\top G^\top \Gamma_{\rm obs}^{-1}\, S_{\rm y}\,)_r = \sum_{i=1}^r \delta_i w_i v_i^\top$, where w_i is the ith column of W, v_i is the ith column of V, and δ_i is the ith diagonal entry of D. Inserting this back into (4.64) yields $A = \sum_{i\le r} \delta_i (1+\delta_i^2)^{-1} S_{\rm pr} w_i v_i^\top S_{\rm obs}^{-1}$. Now it suffices to note that $\widehat{w}_i := S_{\rm pr} w_i$ is a generalized eigenvector of $(H, \Gamma_{\rm pr}^{-1})$,

that $\widehat{v}_i := S_{\rm obs}^{-\top} v_i$ is a generalized eigenvector of $(G\Gamma_{\rm pr}G^\top, \Gamma_{\rm obs})$, and that (δ_i^2) are also eigenvalues of $(H, \Gamma_{\rm pr}^{-1})$. The minimum Bayes risk is a straightforward computation for the optimal estimator (4.28) using (4.60). □

Proof of Theorem 3. Given $A \in \widehat{\mathcal{A}}_r$, we can restate (4.61) as the problem of finding a matrix B, of rank at most r, that minimizes:

$$\| S_{\rm pos}^{-1}(\Gamma_{\rm pr} - \Gamma_{\rm pos})\, G^\top \Gamma_{\rm obs}^{-1}\, S_{\rm y} - S_{\rm pos}^{-1} B \left(G^\top \Gamma_{\rm obs}^{-1}\, S_{\rm y}\right) \|_F \tag{4.65}$$

such that $A = (\Gamma_{\rm pr} - B)\, G^\top \Gamma_{\rm obs}^{-1}$. By [77, Theorem 2.1], an optimal B is given by:

$$B = S_{\rm pos}(\, S_{\rm pos}^{-1}\, (\Gamma_{\rm pr} - \Gamma_{\rm pos})\, G^\top \Gamma_{\rm obs}^{-1}\, S_{\rm y}\,)_r (G^\top \Gamma_{\rm obs}^{-1}\, S_{\rm y})^\dagger \tag{4.66}$$

where $\dagger$ denotes the pseudo-inverse operator. A closer look at [77, Theorem 2.1] reveals that another minimizer of (4.65), itself not necessarily of minimum Frobenius norm, is given by:

$$B = S_{\rm pos}(\, S_{\rm pos}^{-1}\, (\Gamma_{\rm pr} - \Gamma_{\rm pos})\, G^\top \Gamma_{\rm obs}^{-1}\, S_{\rm y}\,)_r\, (S_{\rm pr}^\top G^\top \Gamma_{\rm obs}^{-1}\, S_{\rm y})^\dagger S_{\rm pr}^\top. \tag{4.67}$$

By Lemma 4,

$$\begin{aligned}
S_{\rm pr}^\top G^\top \Gamma_{\rm obs}^{-1}\, S_{\rm y} &= W[\, D \left(I + D^\top D\right)^{1/2}]V^\top \\
S_{\rm pos}^{-1}\, \Gamma_{\rm pr} G^\top \Gamma_{\rm obs}^{-1} S_{\rm y} &= W[\, (I + DD^\top)^{1/2} D (I + D^\top D)^{1/2}]V^\top \\
S_{\rm pos}^{-1}\, \Gamma_{\rm pos} G^\top \Gamma_{\rm obs}^{-1} S_{\rm y} &= W[\, (I + DD^\top)^{-1/2} D (I + D^\top D)^{1/2}]V^\top
\end{aligned}$$

and therefore $(S_{\rm pr}^\top G^\top \Gamma_{\rm obs}^{-1}\, S_{\rm y})^\dagger = \sum_{i=1}^q \delta_i^{-1}\, (1 + \delta_i^2)^{-1/2}\, v_i w_i^\top$ for $q = \mathrm{rank}(S_{\widehat{H}})$, whereas

$$(\, S_{\rm pos}^{-1}\, (\Gamma_{\rm pr} - \Gamma_{\rm pos}) G^\top \Gamma_{\rm obs}^{-1} S_{\rm y}\,)_r = \sum_{i=1}^r \delta_i^3\, w_i v_i^\top.$$

Inserting these expressions back into (4.67), we obtain:

$$B = S_{\text{pr}} \left(\sum_{i=1}^{r} \frac{\delta_i^2}{1+\delta_i^2} \, w_i w_i^\top \right) S_{\text{pr}}^\top,$$

where w_i is the ith column of W, v_i is the ith column of V, and δ_i is the ith diagonal entry of D. Notice that $(\delta_i^2, \widehat{w}_i)$, with $\widehat{w}_i = S_{\text{pr}} w_i$, are the generalized eigenpairs of $(H, \Gamma_{\text{pr}}^{-1})$ (cf. proof of Theorem 1). Hence, by Theorem 1, we recognize the optimal approximation of Γ_{pos} as $\widehat{\Gamma}_{\text{pos}} = \Gamma_{\text{pr}} - B$. Plugging this expression back into (4.67) gives (4.31). The minimum Bayes risk in (ii) follows readily using the optimal estimator given by (4.31) in (4.60). □

Acknowledgments

This work was supported by the US Department of Energy, Office of Advanced Scientific Computing (ASCR), under grant numbers DE-SC0003908 and DE-SC0009297. We thank J. Heikkinen from Bintec Ltd. for providing us with the code used in Example 2.

Chapter 5

A Stochastic Newton MCMC Method for Large-Scale Statistical Inverse Problems with Application to Seismic Inversion

The content of this chapter is based on an existing publication[1] which is joint work with Lucas Wilcox, Carsten Burstedde, and my advisor Omar Ghattas. The forward and adjoint simulations of the physical wave propagation problem were implemented by Carsten, and was completed prior to the beginning of this work. I contributed the majority of the algorithmic implementations and the running of numerical experiments. All authors had significant contribution to the remaining content of this chapter.

Abstract

We address the solution of large-scale statistical inverse problems in the framework of Bayesian inference. The Markov chain Monte Carlo method is the most popular approach for sampling the posterior probability distribution that describes the solution of the statistical inverse problem. MCMC meth-

[1] J. Martin, L. C. Wilcox, C. Burstedde, and O. Ghattas. A stochastic Newton MCMC method for large-scale statistical inverse problems with application to seismic inversion. *SIAM Journal on Scientific Computing*, 34(3):A1460–A1487, 2012. http://epubs.siam.org/doi/abs/10.1137/110845598

ods face two central difficulties when applied to large scale inverse problems: first, the forward models (typically in the form of partial differential equations) that map uncertain parameters to observable quantities make the evaluation of the probability density at any point in parameter space very expensive; and second, the high-dimensional parameter spaces that arise upon discretization of infinite-dimensional parameter fields make the exploration of the pdf prohibitive. The challenge for MCMC methods is to construct proposal functions that simultaneously provide a good approximation of the target density while being inexpensive to manipulate.

Here we present a so-called Stochastic Newton method in which MCMC is accelerated by constructing and sampling from a proposal density that builds a local Gaussian approximation based on local gradient and Hessian (of the log posterior) information. Thus, the method exploits tools (adjoint-based gradients and Hessians) that have been instrumental for fast (often mesh-independent) solution of deterministic inverse problems. Hessian manipulations (inverse, square root) are made tractable by a low rank approximation that exploits the compact nature of the data misfit operator. This is analogous to a reduced model of the parameter-to-observable map. The method is applied to the Bayesian solution of an inverse medium problem governed by 1D seismic wave propagation. We compare the Stochastic Newton method with a reference black box MCMC method as well as a gradient-based Langevin MCMC method, and observe at least two orders of magnitude improvement in convergence for problems with up to 65 parameters. Numerical evidence

suggests that a 1025 parameter problem converges at the same rate as the 65 parameter problem.

5.1 Introduction and background

Uncertainty in reconstructing parameter fields from data is a fundamental feature of ill-posed inverse problems. Our lack of knowledge results from noisy measurements, sparse observations, uncertain forward models, and uncertain prior parameter information. The deterministic output least squares approach to inverse problems, which amounts to minimizing a regularized data misfit function, is incapable of accounting for uncertainties in the solution of the inverse problem. *Bayesian inference* provides a systematic framework for incorporating uncertainties in observations, forward models, and prior knowledge to quantify uncertainties in the model parameters. However, Bayesian solution of large-scale statistical inverse problems, i.e., those described by expensive forward models such as partial differential equations (PDEs), and for large numbers of model parameters as result from discretized parameter fields, is essentially intractable using conventional statistical techniques that view the forward model (i.e., the parameter-to-observable map) as a black box.

We address methods for sampling probability density functions (pdfs) that describe uncertain parameter fields in Bayesian solutions to statistical inverse problems governed by PDEs. Such problems have two properties that present significant challenges for standard Markov chain Monte Carlo (MCMC) sampling methods. First, each sample point requires solution of the forward

problem, which can be exceedingly expensive. Second, discretization of the parameter space can result in very high dimensional pdfs. Here, we present a method that exploits the structure of the inverse operator to greatly speed up MCMC. The method, to which we refer as *Stochastic Newton*, can be derived by analogy with the classical Newton's method for the associated deterministic inverse problem. Stochastic Newton employs a local Gaussian approximation to the target pdf—informed by local Hessian information—as a proposal density for MCMC. A low rank approximation of the Hessian is invoked—reflecting the ill-posed nature of many PDE-based inverse problems—and rendering the computation tractable. Alternatively, Stochastic Newton can be interpreted as a Hessian-preconditioned Langevin MCMC method. In the remainder of this section, we provide background on the Bayesian formulation of statistical inverse problems and on MCMC methods, and discuss alternative approaches.

5.1.1 Bayesian formulation of the statistical inverse problem

The great challenge in solving inverse problems lies in the fact that they are usually ill-posed: many different choices of model parameters may be consistent with the data. Non-uniqueness stems from sparsity of the observations and uncertainty in both the measurements and the model itself. A popular approach to obtaining a unique "solution" to the inverse problem is to formulate it as a least squares optimization problem: minimize the misfit between observed and predicted outputs in an appropriate norm while also minimizing a *regularization* term that penalizes unwanted features of the parameters.

This is often called *Occam's approach*: find the "simplest" set of parameters that is consistent with the measured data. The inverse problem thus leads to a nonlinear optimization problem that is constrained by the forward model. Estimation of parameters using this regularization approach to inverse problems will yield an estimate of the "best" parameter values that simultaneously fit the data and honor the regularization penalty term. However, we are interested in not just point estimates of the best-fit parameters, but a complete statistical description of the parameter values. The *Bayesian* approach does this by reformulating the inverse problem as a problem in *statistical inference*, incorporating uncertainties in the measurements, the forward model, and prior information on the parameters [117, 192]. The solution of this inverse problem is the *posterior* joint probability density of the parameters, which encodes the degree of confidence in their estimate. Thus we are able to quantify the resulting uncertainty in the parameters, taking into account uncertainties in the data, model, and prior information.

Suppose the relationship between output observables $\boldsymbol{d}$ (the predicted outputs at the measurement locations and time instants) and uncertain model parameters $\boldsymbol{m}$ is denoted by $\boldsymbol{d} = f(\boldsymbol{m}, \boldsymbol{e})$, where $\boldsymbol{e}$ represents noise due to measurement and/or modeling errors. In other words, given the model parameters $\boldsymbol{m}$ and noise $\boldsymbol{e}$, the function $f(\boldsymbol{m}, \boldsymbol{e})$ solves the forward (PDE) problem to yield $\boldsymbol{d}$. Suppose also that we have the prior probability density $\pi_{\text{prior}}(\boldsymbol{m})$, which encodes the confidence we have in prior information on the unknown model parameters (i.e., independent of present observations), and the likeli-

hood function $\pi_{\text{like}}(\boldsymbol{d}_{\text{obs}}|\boldsymbol{m})$, which describes the conditional probability that the model parameters $\boldsymbol{m}$ give rise to the actual measurements $\boldsymbol{d}_{\text{obs}}$. Then Bayes' theorem of inverse problems expresses the posterior probability density of the model parameters, π_{post}, given the data $\boldsymbol{d}_{\text{obs}}$, as the conditional probability

$$\pi_{\text{post}}(\boldsymbol{m}) := \pi(\boldsymbol{m}|\boldsymbol{d}_{\text{obs}}) \propto \pi_{\text{prior}}(\boldsymbol{m})\, \pi(\boldsymbol{d}_{\text{obs}}|\boldsymbol{m}). \tag{5.1}$$

Expression (5.1) provides the statistical solution of the inverse problem as a probability density for the model parameters $\boldsymbol{m}$. Often, particularly in high dimensions, we are interested not in a complete characterization of $\pi_{\text{post}}(\boldsymbol{m})$ (which may be intractable to compute and impossible to interpret), but in its moments (mean, covariance, etc.) or other functionals (e.g., event probabilities).

As a specific example, suppose the noise is additive and is modeled as Gaussian with zero mean and a covariance matrix $\boldsymbol{\Gamma}_{\text{noise}}$, and suppose the prior density of the model parameters is represented as Gaussian with $\bar{\boldsymbol{m}}_{\text{prior}}$ as the mean and $\boldsymbol{\Gamma}_{\text{prior}}$ as the covariance matrix, then the posterior probability density of the model parameters is given explicitly (within a normalizing constant) by

$$\pi_{\text{post}}(\boldsymbol{m}) \propto \exp\left[-\tfrac{1}{2}\|f(\boldsymbol{m}) - \boldsymbol{d}_{\text{obs}}\|^2_{\boldsymbol{\Gamma}^{-1}_{\text{noise}}} - \tfrac{1}{2}\|\boldsymbol{m} - \bar{\boldsymbol{m}}_{\text{prior}}\|^2_{\boldsymbol{\Gamma}^{-1}_{\text{prior}}}\right]. \tag{5.2}$$

This latter expression shows that even when the prior, measurement, and modeling uncertainties are Gaussian, the posterior density of the model parameters is generally not Gaussian, due to the nonlinearity of the parameter-to-observable map, $f(\boldsymbol{m})$. However, this expression exposes a significant con-

nection between statistical and deterministic inversion. Suppose we wish to find the value of the most likely model parameters, by maximizing the posterior density (5.2). This is equivalent to minimizing the negative argument of the exponential function—which is precisely the misfit function that is minimized by deterministic inverse methods, provided we interpret the prior as a regularization and weigh the data misfit by the inverse noise covariance. Moreover, it is straightforward to show that the inverse of the Hessian matrix of the deterministic regularized misfit function approximates the covariance matrix of the posterior density (the equivalence is exact when $f(\boldsymbol{m})$ is linear). This connection between the Hessian operator of the deterministic inverse problem and the inverse covariance matrix of the statistical inverse problem is crucial to the computational efficiency of the Stochastic Newton method.

While it is easy to write down expressions for the posterior pdf such as (5.1) or (5.2), making use of these expressions poses a challenge, because the posterior pdf is a surface in high dimensions (equal to the number of model parameters $\boldsymbol{m}$), and because the solution of the forward problem (i.e., computing $f(\boldsymbol{m})$ given $\boldsymbol{m}$) is required to evaluate the probability of any point in parameter space (as can be seen in (5.2)). Straightforward grid-based sampling is limited to problems with a few parameters and cheap forward simulations. Special sampling techniques, such as MCMC methods, have been developed to generate sample ensembles that typically require many fewer points than grid-based sampling, e.g. [117,192,193]. In particular, Metropolis-Hastings (M-H) methods employ a given *proposal* probability density $q(\boldsymbol{m}_k, \boldsymbol{y})$

Algorithm 5 Metropolis-Hastings Algorithm to sample pdf π

Choose initial parameters $\boldsymbol{m}_0$
Compute $\pi(\boldsymbol{m}_0)$
for $k = 0, \ldots, N-1$ **do**
  Draw sample $\boldsymbol{y}$ from the proposal density $q(\boldsymbol{m}_k, \cdot)$
  Compute $\pi(\boldsymbol{y})$
  Compute $\alpha(\boldsymbol{m}_k, \boldsymbol{y}) = \min\left\{1, \frac{\pi(\boldsymbol{y})q(\boldsymbol{y},\boldsymbol{m}_k)}{\pi(\boldsymbol{m}_k)q(\boldsymbol{m}_k,\boldsymbol{y})}\right\}$
  Draw $u \sim \mathcal{U}([0,1))$
  if $u < \alpha(\boldsymbol{m}_k, \boldsymbol{y})$ **then**
    Accept: Set $\boldsymbol{m}_{k+1} = \boldsymbol{y}$
  else
    Reject: Set $\boldsymbol{m}_{k+1} = \boldsymbol{m}_k$
  end if
end for

at each sample point in parameter space $\boldsymbol{m}_k$ to generate a proposed sample point $\boldsymbol{y}$. Once generated, the M-H criterion chooses to either accept or reject the proposed sample point, and repeats from the new point, thereby generating a chain of samples from the posterior density $\pi_{\text{post}}(\boldsymbol{m})$. Algorithm 5.1.1 [117, Section 3.6.2] presents pseudo-code for the M-H method. For example, a popular choice for the proposal density is the isotropic Gaussian $q(\boldsymbol{m}_k, \boldsymbol{y}) = \frac{1}{(2\pi)^{n/2}} \exp[-\frac{1}{2}(\|\boldsymbol{m}_k - \boldsymbol{y}\|)^2]$; the resulting method is known as Random Walk Metropolis. This proposal density is easy to sample, but can lead to poor MCMC performance due to the mismatch between the proposal and posterior densities. This problem is greatly compounded when the parameter dimension is large, and in these cases it is critical that this mismatch is minimized to obtain acceptable MCMC performance. The challenge is to devise a proposal density $q(\boldsymbol{m}_k, \boldsymbol{y})$ that is both easy to sample and a good

representation of the underlying posterior probability density.

A traditional approach is to utilize a single site updating scheme [111, 149]. This approach is more forgiving of naive proposal densities, but requires as many forward simulations as parameters to perform a single parameter sweep. When the forward simulation is expensive and the parameter dimension is large, this approach is computationally intractable. In this paper, we therefore restrict our attention to multivariate proposal densities that update the entire parameter vector at once.

Given the connection between the inverse covariance matrix of the posterior pdf and the Hessian of the deterministic regularized misfit mentioned above, our goal is to capitalize on advances in algorithms for deterministic inverse problems to construct proposal densities for M-H MCMC that exploit the structure of the posterior pdf. In particular, we construct local Gaussian approximations of the posterior pdf from gradient and Hessian information of the negative log posterior. Drawing samples from this proposal density then requires solving systems that are identical to the Newton step for a deterministic inverse problem, thereby exploiting advances in fast Newton methods for deterministic inverse problems. Using modern adjoint techniques, gradients can be computed at a cost of a single linearized forward solve, as can actions of Hessians on vectors. These tools, combined with specialized solvers that exploit the fact that many ill-posed inverse problems have compact data misfit operators, often permit solution of deterministic inverse problems in a dimension-independent (and typically small) number of iterations (e.g., [5]).

We study the efficiency of the Stochastic Newton method introduced here on a model seismic inverse problem, that of recovering the distribution of stiffness of an elastic medium from noisy observations of seismically-induced ground motion at the surface. Stochastic Newton is compared with a freely-available implementation of another popular method that attempts to exploit posterior covariance information, the delayed rejection adaptive Metropolis (DRAM) method, and with unpreconditioned Langevin MCMC. The results demonstrate large speedups over the other methods, and suggest mesh independence of Stochastic Newton for problems with up to 1025 parameters. We demonstrate experimentally that Stochastic Newton is able to take large steps without compromising acceptance rates, and that convergence diagnostics and integrated autocorrelation functions show substantial improvement in the Stochastic Newton sample chains over traditional MCMC sample chains.

5.1.2 Approaches for sampling posterior probability density functions

In this subsection, we review existing approaches to the solution of the statistical inverse problem, and conclude by describing the relationship between the proposed Stochastic Newton method and existing methods. We restrict this review to methods for sampling pdfs that arise specifically from large-scale statistical inverse problems characterized by "expensive" forward models (e.g., those governed by PDEs) and high dimensional parameter spaces (e.g., those that arise by discretization of heterogeneous PDE coefficients). For such problems, nearly every existing method ultimately gives the solution to

the statistical inverse problem as a set of samples drawn from the posterior pdf. To make this sampling tractable, some form of reduction is often advocated. Below we review several different forms of reduction of the forward model and parameter space that have been proposed. We proceed from these reduced modeling approaches to increasingly "intrusive" sampling methods, eventually making use of first and second derivative information to characterize the posterior pdf.

5.1.2.1 Reduced modeling

A popular approach to working with a large number of parameters is to reduce the dimension of the problem in some way during the computation of the (expensive) parameter-to-observable map, and later generate samples by interrogating this reduced representation at a correspondingly reduced cost. Projection-type reduced order models are one possible realization of this idea. Here, the state space is projected onto a limited number of basis functions to obtain an inexpensive reduced forward model. This is then used for posterior evaluation or sampling [10, 34, 78, 153, 200]. In addition, the parameter space may also be reduced to facilitate MCMC methods that work well in low dimensions [136]. The challenge has been to develop reduced models that are faithful over the full high dimensional parameter space.

Other approaches use a truncated polynomial chaos (PC) expansion to represent the uncertain parameters, and construct an approximate stochastic forward problem by Galerkin projection onto this PC basis [83]. This stochas-

tic problem is more expensive than the original forward problem, but once obtained, the solution can be used to construct a surrogate for the posterior distribution, which can be evaluated repeatedly at negligible cost, making it ideal for MCMC sampling [144, 146]. Because the total number of terms in the PC expansion is combinatorial in the parameter dimension, a truncated Karhunen-Loève (KL) expansion (based on the prior distribution) may be employed to prevent the cost of the stochastic forward problem from becoming prohibitive [145]. However, it is necessary to ensure that enough KL modes are retained so that the solution of the statistical inverse problem is not significantly biased toward the prior distribution.

Alternatively, after reduction using a PC basis, one can formulate a functional optimization problem over the stochastic space to characterize the solution to the inverse problem [14]. This idea can be combined with Smolyak sparse grids and stochastic collocation to tackle higher dimensional problems as well [205]. One may also approximate the parameter-to-observable map with a Gaussian process model that is constructed via Bayesian model calibration over a limited set of training data (limited in both the number of experimental observations available, as well as the number of forward model evaluations) [109, 120]. Additionally, the Gaussian process model may incorporate local Hessian information to estimate covariance matrices needed in the Gaussian process representation [32].

5.1.2.2 Adaptive sampling

As an alternative, we may instead "sample then reduce," wherein the full parameter space is sampled by a MCMC method that is able to cope with the high dimensionality and strong correlation structure inherent in ill-posed inverse problems. This can be of particular importance when modes of the parameter space that are important to the inverse problem do not align well with a coordinate basis or strong modes of the prior in the KL expansion, and any reduced basis generated by these approaches would require a prohibitive number of basis vectors to solve the problem with sufficient accuracy.

Delayed Rejection Adaptive Metropolis (DRAM) MCMC adaptively constructs an approximation to the posterior covariance matrix to guide the sampling process and cope with the correlation structure [95]. DRAM requires only the ability to evaluate the posterior density at an arbitrary point, and can thus be considered a black-box (or "non-intrusive") method. Similarly, the so-called t-walk only requires pointwise evaluations, but is specifically designed to be invariant to scale and correlation structure, allowing it to perform well on problems that have different scales or correlations in different regions of parameter space [47].

Many MCMC methods also employ derivative information to help guide sampling, which is more demanding of the types of information that need to be computed from the forward map. Langevin MCMC employs a stochastic differential equation (SDE) that has the desired posterior distribution as a stationary solution. Trajectories (realizations) of this SDE can thus be used

to construct sample chains for the posterior distribution. When discretized, a finite timestep must be selected, and the discrete trajectories may no longer be faithful to the original SDE. Langevin MCMC restores convergence of the sample chain to the desired posterior distribution by considering each timestep as a proposal distribution for the M-H algorithm (see e.g., [8, 180]). This also permits the use of inexpensive approximate gradient information (e.g., computed based on a coarse scale model) [61].

Another class of methods uses a two stage proposal process, where the proposal is first subjected to an accept/reject step based on an inexpensive approximate model (e.g., based on a coarse scale model), and the expensive true solution is computed only when the proposal is likely to be accepted [46, 66, 110].

Finally, Hamiltonian Monte Carlo (HMC) extends the parameter space at each MCMC sample to include a momentum variable, chooses a random sample from momentum space and integrates a Hamiltonian system to generate proposal points. Derivative information of the posterior density is also used for this approach in the construction and solution of this system. A review of HMC methods can be found in [152].

5.1.2.3 Hessian-based sampling

Last but not least, we consider methods that make use of Hessian information (i.e., second derivatives) of the forward map. This information is generally more expensive to obtain, but can prove highly beneficial to speeding

up convergence of the sampling process. MCMC methods that utilize Hessian information have been considered previously [81, 82, 171], but are practicable only for a small number of parameters or for problems where an analytical expression for the Hessian is available. In [103], a BFGS-type approximation of the Hessian is considered for this purpose to avoid explicit computation of second derivatives, and demonstrated on a 16 parameter Gaussian posterior distribution.

Another interesting approach makes use of the Fisher information as a natural metric for a Riemannian manifold [86]. Langevin MCMC and HMC can both be derived in this particular metric, and show significant gains over the traditional varieties of MCMC by respecting the local structure of the parameter space. This method employs what amounts to the Gauss-Newton approximation of the Hessian of the negative log posterior, as well as additional third derivative terms. Computing the exact Gauss-Newton Hessian is generally intractable for large-scale inverse problems since it requires the solution of as many forward problems as the number of parameters.

Finally, the Stochastic Newton method we introduce in this paper can be understood as a relative of a preconditioned Langevin MCMC method, where the preconditioning is performed with the local Hessian of the negative log posterior. It is noteworthy that we obtain a similar preconditioning term to the one that appears in the Riemannian-manifold derivation of Langevin MCMC; in this paper however, we construct an accurate low-rank representation of the Hessian, and show that all necessary computations can

be performed without constructing the full Hessian operator. This permits scalability to large parameters dimensions.

5.1.3 Outline of the paper

In Section 5.2, we demonstrate the natural connections between deterministic optimization and the statistical inverse problem, use these connections to derive Stochastic Newton MCMC, and derive the low-rank approximations required to make the method tractable for large-scale inverse problems. Section 5.3 presents a motivating Bayesian statistical inverse problem based on seismic wave scattering. Finally, in Section 5.4 we compare the performance of Stochastic Newton with DRAM MCMC and Langevin MCMC in various convergence metrics, demonstrate that Stochastic Newton offers a favorable trade off between increased complexity of the computations and improved MCMC convergence, and show examples that support good scalability with increasing dimensionality of parameter space.

5.2 Stochastic Newton MCMC

Large scale optimization provides many tools and insights—in particular, Newton's method and its matrix-free variants—that accelerate the solution of deterministic inverse problems. In this section, we develop the Stochastic Newton method, which exploits natural connections between the deterministic inverse problem and the Bayesian statistical inverse problem to accelerate statistical sampling methods. Moreover, motivated by the spectral structure

of underlying infinite dimensional Hessian operators that appear in many ill-posed inverse problems, we introduce low rank approximations that make the Stochastic Newton method tractable in high dimensions.

5.2.1 Connection with optimization

Consider a finite dimensional parameter-to-observable map $\boldsymbol{d} = \boldsymbol{f}(\boldsymbol{m})$ that maps parameters $\boldsymbol{m} \in \mathbb{R}^n$ to observables $\boldsymbol{d} \in \mathbb{R}^m$. The deterministic inverse problem seeks to minimize $\frac{1}{2}\|\boldsymbol{f}(\boldsymbol{m}) - \boldsymbol{d}_{\text{obs}}\|^2_{\boldsymbol{W}}$, the misfit between the model predictions and the observed data $\boldsymbol{d}_{\text{obs}} \in \mathbb{R}^m$ in the $\boldsymbol{W}$-norm, with $\boldsymbol{W} \in \mathbb{R}^m \times \mathbb{R}^m$. A quadratic regularization term $\frac{1}{2}\|\boldsymbol{m} - \bar{\boldsymbol{m}}\|^2_{\boldsymbol{R}}$ penalizes distance from a baseline vector of parameters $\bar{\boldsymbol{m}} \in \mathbb{R}^n$ in the $\boldsymbol{R}$-norm, with $\boldsymbol{R} \in \mathbb{R}^n \times \mathbb{R}^n$. Appropriate regularization of this form addresses ill-posedness of the inverse problem and guarantees uniqueness of the solution $\boldsymbol{m}^*$ to the following deterministic inverse problem:

$$\boldsymbol{m}^* = \arg\min_{\boldsymbol{m}} \left(\frac{1}{2}\|\boldsymbol{f}(\boldsymbol{m}) - \boldsymbol{d}_{\text{obs}}\|^2_{\boldsymbol{W}} + \frac{1}{2}\|\boldsymbol{m} - \bar{\boldsymbol{m}}\|^2_{\boldsymbol{R}} \right), \tag{5.3}$$

In the statistical inverse setting, we observe that Bayes' Theorem (5.1) can be understood directly in the deterministic context if we inspect the negative log-posterior,

$$-\log \pi_{\text{post}} = -\log \pi_{\text{like}} - \log \pi_{\text{prior}} + \text{const.} \tag{5.4}$$

The constant of proportionality from Bayes' Theorem is included above, but affects neither the deterministic optimization nor the statistical inverse problem. In the statistical setting of the inverse problem, the misfit $\boldsymbol{f}(\boldsymbol{m}) - \boldsymbol{d}_{\text{obs}}$ is

interpreted as a vector-valued random variable. When the measurement error and model error are unbiased, additive, and Gaussian, we have $(\boldsymbol{f}(\boldsymbol{m})-\boldsymbol{d}_{\text{obs}}) \sim \mathcal{N}(\mathbf{0}, \boldsymbol{\Gamma}_{\text{noise}})$. (See e.g. Section 3.2.1 of [117].) The log-likelihood function in this context plays the role of the misfit term in the optimization formulation (5.3):

$$-\log \pi_{\text{like}}(\boldsymbol{d}_{\text{obs}}|\boldsymbol{m}) = \tfrac{1}{2}(\boldsymbol{f}(\boldsymbol{m}) - \boldsymbol{d}_{\text{obs}})^T \boldsymbol{\Gamma}_{\text{noise}}^{-1}(\boldsymbol{f}(\boldsymbol{m}) - \boldsymbol{d}_{\text{obs}}). \tag{5.5}$$

More general considerations of measurement and model error are possible, and do not restrict the applicability of our method.

Similarly, if the prior density is Gaussian with mean $\bar{\boldsymbol{m}}_{\text{prior}}$ and covariance matrix $\boldsymbol{\Gamma}_{\text{prior}}$, then the log-prior term in (5.4) plays the role of the regularization from deterministic optimization,

$$-\log \pi_{\text{prior}}(\boldsymbol{m}) = \tfrac{1}{2}(\boldsymbol{m} - \bar{\boldsymbol{m}}_{\text{prior}})^T \boldsymbol{\Gamma}_{\text{prior}}^{-1}(\boldsymbol{m} - \bar{\boldsymbol{m}}_{\text{prior}}). \tag{5.6}$$

As before the assumption of a Gaussian prior can also be relaxed.

The negative log-posterior (5.4) is now understood directly as the cost function $V(\boldsymbol{m})$ from deterministic optimization, and therefore we can write the posterior density as

$$\pi_{\text{post}}(\boldsymbol{m}|\boldsymbol{d}_{\text{obs}}) \propto \exp(-V(\boldsymbol{m})), \tag{5.7}$$

where the cost function $V(\boldsymbol{m})$ is given by

$$V(\boldsymbol{m}) := \tfrac{1}{2}\|f(\boldsymbol{m}) - \boldsymbol{d}_{\text{obs}}\|^2_{\boldsymbol{\Gamma}_{\text{noise}}^{-1}} + \tfrac{1}{2}\|\boldsymbol{m} - \bar{\boldsymbol{m}}_{\text{prior}}\|^2_{\boldsymbol{\Gamma}_{\text{prior}}^{-1}}. \tag{5.8}$$

The connection between the cost functional from deterministic optimization and the posterior probability density in the statistical setting is made explicit when we seek the maximum *a posteriori* (MAP) estimate $\boldsymbol{m}_{\text{MAP}}$, which is given by maximizing the posterior, or equivalently, by minimizing the cost function. Thus, $\boldsymbol{m}_{\text{MAP}} = \boldsymbol{m}^*$ when the appropriate definitions of $\boldsymbol{W}$ and $\boldsymbol{R}$ are taken in (5.3). Next, we consider how to further exploit this connection between deterministic and statistical inversion.

5.2.2 The Gaussian linear case

When the parameter-to-observable map is linear, we write $\boldsymbol{f}(\boldsymbol{m}) = \boldsymbol{G}\boldsymbol{m}$ with $\boldsymbol{G} \in \mathbb{R}^{m\times n}$. In this case we observe that the negative log posterior (or deterministic cost function)

$$V(\boldsymbol{m}) = \tfrac{1}{2}(\boldsymbol{G}\boldsymbol{m}-\boldsymbol{d}_{\text{obs}})^T\boldsymbol{\Gamma}_{\text{noise}}^{-1}(\boldsymbol{G}\boldsymbol{m}-\boldsymbol{d}_{\text{obs}}) + \tfrac{1}{2}(\boldsymbol{m}-\bar{\boldsymbol{m}}_{\text{prior}})^T\boldsymbol{\Gamma}_{\text{prior}}^{-1}(\boldsymbol{m}-\bar{\boldsymbol{m}}_{\text{prior}}) \tag{5.9}$$

is quadratic in the parameters $\boldsymbol{m}$. Making use of the expressions for the gradient $\boldsymbol{g}$ and Hessian $\boldsymbol{H}$ of $V(\boldsymbol{m})$,

$$\boldsymbol{g} := \boldsymbol{g}(\boldsymbol{m}) = \nabla V(\boldsymbol{m}) = \boldsymbol{G}^T\boldsymbol{\Gamma}_{\text{noise}}^{-1}(\boldsymbol{G}\boldsymbol{m}-\boldsymbol{d}_{\text{obs}}) + \boldsymbol{\Gamma}_{\text{prior}}^{-1}(\boldsymbol{m}-\bar{\boldsymbol{m}}_{\text{prior}}), \tag{5.10}$$

$$\boldsymbol{H} := \nabla^2 V(\boldsymbol{m}) = \boldsymbol{G}^T\boldsymbol{\Gamma}_{\text{noise}}^{-1}\boldsymbol{G} + \boldsymbol{\Gamma}_{\text{prior}}^{-1}, \tag{5.11}$$

we can rewrite the cost function in the form

$$V(\boldsymbol{m}) = \tfrac{1}{2}(\boldsymbol{m}-\boldsymbol{m}^*)^T\boldsymbol{H}(\boldsymbol{m}-\boldsymbol{m}^*) + \text{const}, \tag{5.12}$$

which makes clear, since both $\boldsymbol{\Gamma}_{\text{noise}}$ and $\boldsymbol{\Gamma}_{\text{prior}}$ and therefore $\boldsymbol{H}$ are positive definite, that a unique minimum of $V(\boldsymbol{m})$ exists and is given by requiring

$\boldsymbol{g}(\boldsymbol{m}^*) = \boldsymbol{0}$,

$$\boldsymbol{m}^* = \boldsymbol{H}^{-1}\left(\boldsymbol{G}^T\boldsymbol{\Gamma}_{\text{noise}}^{-1}\boldsymbol{d}_{\text{obs}} + \boldsymbol{\Gamma}_{\text{prior}}^{-1}\bar{\boldsymbol{m}}_{\text{prior}}\right). \tag{5.13}$$

Moreover, the posterior pdf $\exp(-V(\boldsymbol{m}))$ can be seen to be Gaussian with mean given by the minimizer of $V(\boldsymbol{m})$, i.e. the solution of the deterministic inverse problem (5.3), and covariance given by the inverse of the Hessian, $\boldsymbol{H}^{-1}$, i.e., π_{post} is distributed as $\mathcal{N}(\boldsymbol{m}^*, \boldsymbol{H}^{-1})$. Hence, we see an explicit connection between the deterministic solution and its statistical counterparts, at least in the case of a linear parameter-to-observable map.

5.2.3 The nonlinear case and Stochastic Newton's method

When the parameter-to-observable map $\boldsymbol{f}(\boldsymbol{m})$ is nonlinear, the posterior is no longer Gaussian, and in general the minimum of the cost function no longer coincides with the mean of the posterior, nor does the inverse of the Hessian coincide with the covariance matrix of the posterior. However, we can still exploit connections between deterministic optimization methods for minimizing $V(\boldsymbol{m})$ and statistical methods for sampling the posterior π_{post}.

The gold standard for optimization is Newton's method, which begins with a local quadratic approximation $\tilde{V}(\boldsymbol{m})$ of the cost function about a given point $\boldsymbol{m}_k$, which can be written as

$$V(\boldsymbol{m}) \approx \tilde{V}(\boldsymbol{m}) = \frac{1}{2}(\boldsymbol{m}-\boldsymbol{m}_k)^T\boldsymbol{H}(\boldsymbol{m}-\boldsymbol{m}_k) + \boldsymbol{g}^T(\boldsymbol{m}-\boldsymbol{m}_k) + V(\boldsymbol{m}_k), \tag{5.14}$$

with gradient $\boldsymbol{g}(\boldsymbol{m}_k) = \nabla V(\boldsymbol{m}_k)$ and Hessian $\boldsymbol{H}(\boldsymbol{m}_k) = \nabla^2 V(\boldsymbol{m}_k)$.

In the vicinity of a local minimum, $\boldsymbol{H}$ is positive definite. However, at

an arbitrary point $\boldsymbol{m}$, $\boldsymbol{H}$ is not guaranteed to be positive definite, and in such cases it is necessary to replace $\boldsymbol{H}$ with a suitably modified positive definite Hessian $\tilde{\boldsymbol{H}}$ in $\tilde{V}(\boldsymbol{m})$ in order to guarantee convergence. A simple choice for $\tilde{\boldsymbol{H}}$ is an eigenvalue decomposition of $\boldsymbol{H}$, with small or negative eigenvalues replaced with a minimum threshold value. Finally, we rearrange (5.14) as we did in the Gaussian linear case (5.12) to write

$$\tilde{V}(\boldsymbol{m}) = \frac{1}{2}(\boldsymbol{m} - \boldsymbol{m}_k + \tilde{\boldsymbol{H}}^{-1}\boldsymbol{g})^T \tilde{\boldsymbol{H}}(\boldsymbol{m} - \boldsymbol{m}_k + \tilde{\boldsymbol{H}}^{-1}\boldsymbol{g}) + \text{const}, \tag{5.15}$$

which shows that the minimizer of $\tilde{V}(\boldsymbol{m})$ is given by $\boldsymbol{m}_{k+1} = \boldsymbol{m}_k - \tilde{\boldsymbol{H}}^{-1}\boldsymbol{g}$. Note that $-\tilde{\boldsymbol{H}}^{-1}\boldsymbol{g}$ is the Newton step, and iterating this process leads to the classical Newton's method.

In the statistical setting, inserting $\tilde{V}(\boldsymbol{m})$ into (5.7) leads to an approximation of π_{post} given by

$$\pi_{\text{post}}(\boldsymbol{m}) \approx \tilde{\pi}(\boldsymbol{m}) = \exp(-\tilde{V}(\boldsymbol{m})), \tag{5.16}$$

which is in fact a Gaussian, centered on the point $\boldsymbol{m}_{k+1}$ (the result of the the deterministic Newton step) with covariance $\tilde{\boldsymbol{H}}^{-1}$.

Having constructed a local Gaussian approximation of the posterior pdf, we are now in a position to define the Stochastic Newton method, which is a MCMC method that uses the normalized proposal density

$$\tilde{\pi}(\boldsymbol{y}) = \frac{\det \tilde{\boldsymbol{H}}^{1/2}}{(2\pi)^{n/2}} \exp\left(-\frac{1}{2}\left(\boldsymbol{y} - \boldsymbol{m}_k + \tilde{\boldsymbol{H}}^{-1}\boldsymbol{g}\right)^T \tilde{\boldsymbol{H}}\left(\boldsymbol{y} - \boldsymbol{m}_k + \tilde{\boldsymbol{H}}^{-1}\boldsymbol{g}\right)\right). \tag{5.17}$$

Recall that the quadratic approximation is constructed using gradient and Hessian information at $\boldsymbol{m}_k$, namely $\tilde{\boldsymbol{H}}(\boldsymbol{m}_k)$ and $\boldsymbol{g}(\boldsymbol{m}_k)$. Thus, we have "tailored" the proposal density $q(\boldsymbol{m}_k, \boldsymbol{y}) = \tilde{\pi}(\boldsymbol{y})$ to the underlying posterior pdf using derivative information of $V(\boldsymbol{m})$. The Stochastic Newton step at each MCMC iteration proposes a sample $\boldsymbol{y}$ from the density $\tilde{\pi}(\boldsymbol{y})$, which is then subjected to the accept/reject framework of the Metropolis-Hastings algorithm. Pseudocode for Stochastic Newton MCMC for this problem is given in Algorithm 6.

If in fact the posterior density π_{post} is Gaussian (e.g., Section 5.2.2), and the Hessian $\tilde{\boldsymbol{H}}$ is exact, then $q(\boldsymbol{m}_k, \boldsymbol{y}) = \tilde{\pi}(\boldsymbol{y}) = \pi_{\text{post}}(\boldsymbol{y})$, and the Metropolis-Hastings acceptance probability in Algorithm 6 reduces to

$$\alpha(\boldsymbol{m}_k, \boldsymbol{y}) = \min\left\{1, \frac{\pi(\boldsymbol{y})q(\boldsymbol{y}, \boldsymbol{m}_k)}{\pi(\boldsymbol{m}_k)q(\boldsymbol{m}_k, \boldsymbol{y})}\right\} = \min\left\{1, \frac{\pi(\boldsymbol{y})\pi(\boldsymbol{m}_k)}{\pi(\boldsymbol{m}_k)\pi(\boldsymbol{y})}\right\} = 1. \quad (5.18)$$

Thus in this case we achieve "perfect sampling," in which all samples are independent draws from the true posterior density $\pi_{\text{post}}(\boldsymbol{m})$, and accepted with probability 1.

Before concluding this section, we make one final remark about the threshold value used to define $\tilde{\boldsymbol{H}}$. Because $\tilde{\boldsymbol{H}}^{-1}$ is used as the covariance matrix for the proposal distribution, this minimum threshold value for $\tilde{\boldsymbol{H}}$ guarantees a maximum covariance value for the proposal density. This threshold value can therefore be used as a tunable parameter in MCMC to restrict the maximum desired step length to improve performance if the sample acceptance rate is too low.

Algorithm 6 Stochastic Newton MCMC Algorithm to sample π_{post}

Choose initial $\boldsymbol{m}_0$
Compute $\pi_{\text{post}}(\boldsymbol{m}_0), \boldsymbol{g}(\boldsymbol{m}_0), \boldsymbol{H}(\boldsymbol{m}_0)$
for $k = 0, \ldots, N-1$ **do**
  Define $q(\boldsymbol{m}_k, \boldsymbol{y}) = \tilde{\pi}(\boldsymbol{y})$ as in equation (5.17)
  Draw sample $\boldsymbol{y}$ from the proposal density $q(\boldsymbol{m}_k, \cdot)$
  Compute $\pi_{\text{post}}(\boldsymbol{y}), \boldsymbol{g}(\boldsymbol{y}), \boldsymbol{H}(\boldsymbol{y})$
  Compute $\alpha(\boldsymbol{m}_k, \boldsymbol{y}) = \min\left(1, \frac{\pi_{\text{post}}(\boldsymbol{y}) q(\boldsymbol{y}, \boldsymbol{m}_k)}{\pi_{\text{post}}(\boldsymbol{m}_k) q(\boldsymbol{m}_k, \boldsymbol{y})}\right)$
  Draw $u \sim \mathcal{U}([0,1])$
  if $u < \alpha(\boldsymbol{m}_k, \boldsymbol{y})$ **then**
    Accept: Set $\boldsymbol{m}_{k+1} = \boldsymbol{y}$
  else
    Reject: Set $\boldsymbol{m}_{k+1} = \boldsymbol{m}_k$
  end if
end for

5.2.4 Low-rank Hessian approximation

The MCMC method we are proposing here has been contemplated before [82, 171, 174], but applied only to low-dimensional sampling problems and not computationally-intensive inverse problems as we consider here. Attempting to apply the method as described above to such problems will quickly lead to failure, since constructing just one Hessian requires n forward solves (e.g., [21, 112]), that is, equal to the number of parameters. Thus MCMC becomes intractable for expensive forward problems (e.g., governed by PDEs) and in high dimensions (e.g., when the parameters describe a discretization of a field, such as a PDE coefficient, initial condition, boundary condition, etc.).

However, experience with large-scale deterministic inverse problems has shown in many cases that the Hessian of the data misfit term in (5.3) is a

compact operator whose range space is independent of mesh resolution (see e.g., [198]). The intuition behind the compactness of the Hessian of the data misfit term,

$$\boldsymbol{H}_{\text{misfit}} = -\nabla^2 \log \pi_{\text{like}}, \tag{5.19}$$

is that for many ill-posed inverse problems, the observations are sparse and typically inform only a limited number of modes of the parameter field; thus, the Jacobian matrix of observables $\boldsymbol{f}(\boldsymbol{m})$ with respect to parameters $\boldsymbol{m}$ is well-approximated by a low-rank matrix. In particular, it can be shown that the Hessian of the data misfit operator for the inverse medium scattering problem we consider in §5.3 is a compact operator with exponentially decaying spectrum (when the medium is analytic) [30]. This property suggests a low rank approximation of the data misfit Hessian, which permits us to avoid prohibitive computation of the full Hessian. Below, we exploit the compactness of the data misfit Hessian to make the Stochastic Newton MCMC method presented here tractable for large-scale problems.

In the Bayesian setting, the Hessian $\boldsymbol{H}$ can be written as a sum of data misfit and prior Hessians, i.e.,

$$\boldsymbol{H} = \boldsymbol{H}_{\text{misfit}} + \boldsymbol{\Gamma}_{\text{prior}}^{-1}. \tag{5.20}$$

Consider a decomposition of the prior such that $\boldsymbol{\Gamma}_{\text{prior}} = \boldsymbol{L}\boldsymbol{L}^T$, computed either as the symmetric square root $\boldsymbol{L} = \boldsymbol{\Gamma}_{\text{prior}}^{1/2}$, or as the Cholesky factorization.[2]

[2] For problems with very large parameter dimension, this factorization may become

Rewriting $\boldsymbol{H}$ as

$$\boldsymbol{H} = \boldsymbol{L}^{-T}\left(\boldsymbol{L}^T\boldsymbol{H}_{\text{misfit}}\boldsymbol{L} + \boldsymbol{I}\right)\boldsymbol{L}^{-1}, \tag{5.21}$$

we see that the expression $\boldsymbol{L}^T\boldsymbol{H}_{\text{misfit}}\boldsymbol{L}$ emerges as a natural candidate for a low rank spectral approximation, since comparison with the identity provides a quantitative criterion for truncating the spectrum, and since $\boldsymbol{\Gamma}_{\text{prior}}$ is often a smoothing operator, and thus the collapse of the spectrum of $\boldsymbol{H}_{\text{misfit}}$ is then enhanced by preconditioning with $\boldsymbol{L}$. The low rank approximation of $\boldsymbol{L}^T\boldsymbol{H}_{\text{misfit}}\boldsymbol{L}$ represents the parameter subspace in which the data are most informative about the parameters, and least constrained by the prior.

Using Lanczos (or any of its siblings [183]), an r-dimensional low rank approximation can be represented as $\boldsymbol{L}^T\boldsymbol{H}_{\text{misfit}}\boldsymbol{L} \approx \boldsymbol{V}_r\boldsymbol{D}_r\boldsymbol{V}_r^T$, where $\boldsymbol{V}_r \in \mathbb{R}^{n\times r}$ contains the dominant eigenvectors, and the diagonal matrix $\boldsymbol{D}_r \in \mathbb{R}^{r\times r}$ contains the dominant eigenvalues. The number of desired eigenvectors r is determined by truncating the decomposition once the eigenvalues fall below some threshold value $\alpha \ll 1$, below which it is assumed that the prior dominates the character of the Hessian. This truncation also ensures positive definiteness of the low rank approximation of $\boldsymbol{H}$, which we identify as $\tilde{\boldsymbol{H}}$.

Tractability of the low rank approximation and its use in the Stochastic Newton Method can be established as follows. First, Lanczos requires only

prohibitively costly to perform. In this case, one prefers to exploit the structure of the prior to specify $\boldsymbol{L}$ or $\boldsymbol{\Gamma}_{\text{prior}}^{1/2}$ directly—or their action on a vector—via an $O(n)$ method. For example, for a smoothing prior, one can view $\boldsymbol{\Gamma}_{\text{prior}}^{1/2}$ as an elliptic solve via a multigrid method [26, 189].

matrix-vector products ("matvecs") and therefore there is no need to explicitly form the (dense) Hessian. Second, Lanczos tends to perform only as many matvecs as there are extreme (dominant) eigenvalues, so that compactness of the Hessian bounds the number of required Lanczos iterations. Third, each matvec requires only a pair of forward and adjoint PDE solves (e.g., [21], [112, §1.6.5]). Therefore, the approximation can be constructed in a number of PDE solves comparable to the number of dominant eigenvalues, r. For many ill-posed inverse problems in which the parameters are a discretization of an unknown field, the dominant eigenvalues are associated with smooth eigenvectors (physically, this is a consequence of the data being uninformative about small length scales); as such, the dominant eigenvalues are unaffected by subsequent refinement, once a suitable discretization level is achieved. Thus, r is often independent of n (see, for example, [26]). Finally, we observe that all necessary MCMC computations involving the Hessian can be performed without ever explicitly constructing the dense operator, as follows:

$$\tilde{\boldsymbol{H}} = \boldsymbol{L}^{-T}\Big[\boldsymbol{V}_r\boldsymbol{D}_r\boldsymbol{V}_r^T + \boldsymbol{I}\Big]\boldsymbol{L}^{-1} \tag{5.22}$$

$$\tilde{\boldsymbol{H}}^{-1}\boldsymbol{g} = \boldsymbol{L}\Big\{\boldsymbol{V}_r\Big[(\boldsymbol{D}_r + \boldsymbol{I}_r)^{-1} - \boldsymbol{I}_r\Big]\boldsymbol{V}_r^T + \boldsymbol{I}\Big\}\boldsymbol{L}^T\boldsymbol{g} \tag{5.23}$$

$$\tilde{\boldsymbol{H}}^{-1/2}\boldsymbol{x} = \boldsymbol{L}\Big\{\boldsymbol{V}_r\Big[(\boldsymbol{D}_r + \boldsymbol{I}_r)^{-1/2} - \boldsymbol{I}_r\Big]\boldsymbol{V}_r^T + \boldsymbol{I}\Big\}\boldsymbol{x} \tag{5.24}$$

$$\det(\tilde{\boldsymbol{H}}^{1/2}) = (\det\boldsymbol{L})^{-1}\prod_{i=1}^{r}(d_i+1)^{1/2} \tag{5.25}$$

Expression (5.23) computes the Newton step, (5.24) allows us to sample from a Gaussian distribution with covariance $\tilde{\boldsymbol{H}}^{-1}$, and finally (5.25) is necessary in the computation of the accept/reject criterion of Metropolis-Hastings. With

the exception of operations with the square root of the prior, $\boldsymbol{L}$, the complexity of operations in (5.23)–(5.24) is $O(rn)$, where as noted above r is often independent of n. The determinant (5.25) requires only $O(r)$ operations in practice, since $\det(\boldsymbol{L})$ can be precomputed once. Finally, the complexity of carrying out operations with $\boldsymbol{L}$ in (5.23)–(5.24) appears naively to be $O(n^2)$; however, as mentioned in Footnote 1, for very large scale problems (particularly on parallel computers), one would avoid a naive factorization, and instead interpret the action of $\boldsymbol{L}$ on a vector using a fast solver [26].

In summary, the low rank representation $\tilde{\boldsymbol{H}}$ can in many cases be computed efficiently (i.e., in a small number of PDE solves, independent of mesh, and therefore of problem, size), and applied in $O(n)$ computational work.

5.2.5 Comparison with Langevin MCMC Methods

Stochastic Newton also has a natural interpretation as a form of a Langevin MCMC method. In Langevin MCMC, we begin with the negative log posterior $V(\boldsymbol{m})$, and construct trajectories of the stochastic process from Langevin dynamics,

$$d\boldsymbol{X}_t = -\boldsymbol{A}\nabla V \mathrm{d}t + \sqrt{2}\boldsymbol{A}^{1/2} d\boldsymbol{W}_t, \tag{5.26}$$

which sample the desired probability density as $t \to \infty$ [190]. Here, $\boldsymbol{A}$ is a positive definite preconditioning matrix, and stochastic variables are denoted by $\boldsymbol{X}_t$ and $\boldsymbol{W}_t$, where $\boldsymbol{W}_t$ is the vector of standard independent Brownian Motions. When $\boldsymbol{A}$ is the identity, we recover traditional Langevin dynamics.

To solve (5.26), time is discretized by the Euler Maruyama method, with time step Δt, to yield the update

$$\boldsymbol{x}_{k+1} - \boldsymbol{x}_k = -\boldsymbol{A}\nabla V(\boldsymbol{x}_k)\Delta t + \mathcal{N}(\boldsymbol{0}, 2\Delta t\boldsymbol{A}). \tag{5.27}$$

Preconditioning with the local inverse Hessian $\boldsymbol{H}^{-1}$, choosing $\Delta t = 1$, and discarding the factor of 2, we can formally recover the Stochastic Newton's method derived previously,

$$\boldsymbol{x}_{k+1} - \boldsymbol{x}_k = -\boldsymbol{H}^{-1}\nabla V(\boldsymbol{x}_k) + \mathcal{N}(\boldsymbol{0}, \boldsymbol{H}^{-1}). \tag{5.28}$$

Of course the Hessian is both nonconstant and not everywhere positive definite, and so Stochastic Newton is not rigorously understood as a Langevin MCMC, but there do exist definite parallels. Note that without preconditioning, i.e.,

$$\boldsymbol{x}_{k+1} - \boldsymbol{x}_k = -\nabla V(\boldsymbol{x}_k)\Delta t + \mathcal{N}(\boldsymbol{0}, 2\Delta t\boldsymbol{I}), \tag{5.29}$$

Langevin MCMC resembles a steepest descent method in the deterministic setting.

5.2.6 Comparison to other Gaussian MCMC proposal types

Stochastic Newton's use of a Hessian-based local Gaussian approximation as a proposal function can be contrasted with other types of Gaussian proposal functions. Figure 5.1 shows proposal density contours for several different proposal functions, using the Rosenbrock function as an example target density. All contours in the image are normalized so that they contain 5%, 50%, and 95% of the density respectively. In this way, the best acceptance

rates and sample chain convergence will be achieved for the proposal that matches the contours of the target density most closely.

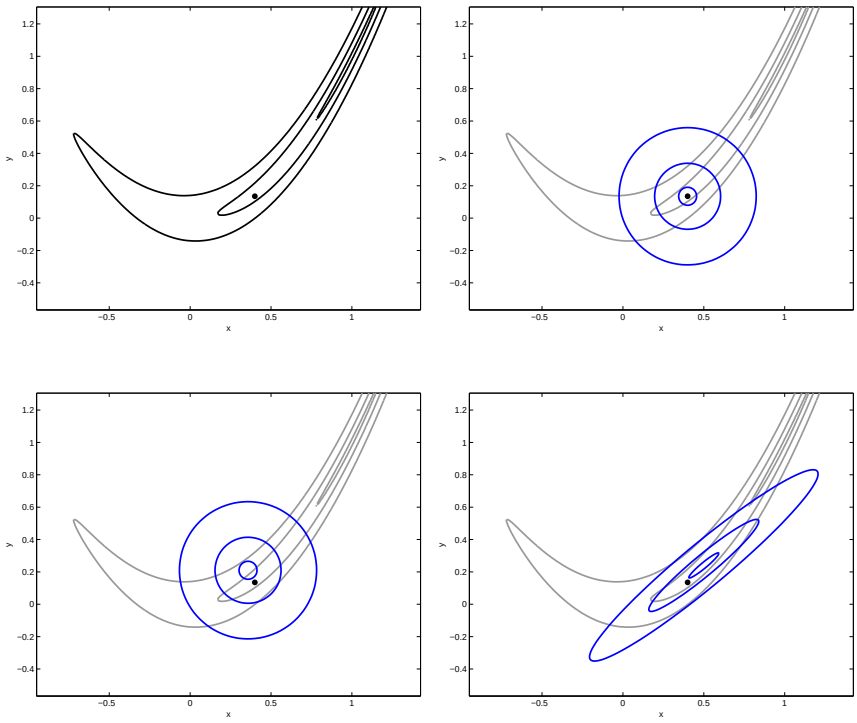

Figure 5.1: Visualizations of differing types of proposal distributions for MCMC. Top left shows contours of the classical Rosenbrock function from deterministic optimization, with effective (unnormalized) density $\exp\Big(-(1-x)^2-100(y-x^2)^2\Big)$. Top right shows the contours of the random walk proposal function overlaid on the Rosenbrock contours. Bottom left shows overlays of Langevin contours without preconditioning. Bottom right shows contours of the Stochastic Newton's Method type proposal function.

5.3 Application to statistical seismic inverse problem

We demonstrate the Stochastic Newton's Method by solving a particular statistical inverse problem. Consider a theoretical seismic exploration

experiment in which a surface explosion causes seismic waves to travel downward into the subsurface medium. If there are obstacles in the medium, or if the medium properties vary with depth, then a fraction of the seismic wave energy will scatter off of these boundaries and return to the surface to be observed at later times. The statistical inverse problem processes these observations to reconstruct a statistical description of the subsurface medium properties. Using this description we are able to estimate properties of the subsurface, including locations of buried objects or oil/mineral deposits.

The remainder of this section describes in detail the ingredients required for this statistical reconstruction. The first ingredient is the mathematical forward model, which maps input medium parameters to predicted observations. Second, the likelihood function uses these predicted observations to determine the probability that the given input medium parameters could have produced the observed experimental data. The third ingredient in any Bayesian analysis is the prior distribution, which encapsulates all of the assumptions about the subsurface medium before any experimental data is considered. Bayes' theorem combines these ingredients into the posterior probability distribution over the set of input medium parameters, which is the statistical description of the subsurface medium which we seek. Finally, we describe the efficient computation of adjoint, gradient and Hessian-vector product information which is required for use of our method.

5.3.1 The forward model

We model our exploration experiment using the 1-D wave equation. The problem is solved on the spatial domain $\Omega = [0, L]$, where $z \in \Omega$ represents the depth beneath the surface at $z = 0$. At the maximum depth $z = L$, we use an absorbing boundary condition which allows plane waves to pass through the boundary without reflection.

The surface explosion is modeled with a right hand side forcing input to the wave equation using a Ricker wavelet $F^{\text{ricker}}(t)$ in time with a mean spectrum energy density at 0.5 Hertz, and a spatial delta function at the surface $\delta(z - 0)$.

Finally, our model has two physical parameters, which are the density ρ and a stiffness parameter μ. In principle both parameters may vary freely with depth, but we will only consider variations in the stiffness $\mu(z; \boldsymbol{m})$, and assume a constant density $\rho = 1$. Note that we have included an explicit dependence on the model parameters $\boldsymbol{m}$.

The governing equations for the forward model are:

$$\rho\, u_{tt}(z,t) - \Big(\mu(z;\boldsymbol{m})\, u_z(z,t)\Big)_z = F^{\text{ricker}}(t)\delta(z-0) \qquad \text{(PDE)}$$

$$\mu(L;\boldsymbol{m})\, u_z(L,t) = -\sqrt{\rho\,\mu(L;\boldsymbol{m})}\, u_t(L,t) \qquad \text{(Absorbing BC)}$$

$$\mu(0;\boldsymbol{m})\, u_z(0,t) = 0 \qquad \text{(Free Surface BC)}$$

$$u(z,0) = 0 \qquad \text{(IC)}$$

$$u_t(z,0) = 0 \qquad \text{(IC)}$$

These equations are solved numerically using Finite Elements on piecewise linear meshes in space using an explicit scheme in time, as in [35]. Most of the examples here are solved on 64 element (65 DOF) meshes, and a few are solved on 1024 element (1025 DOF) meshes. The physical parameter $\mu(z;\boldsymbol{m})$ is discretized as a linear combination of the same 65 or 1025 degrees of freedom as the numerical PDE solution.

Finally, we observe the system by measuring the surface displacement at 120 equally spaced points in time. These measurements are assumed to contain errors at each observation time which are Gaussian, additive, and independent. The noise level is selected such that the resulting RMS signal to noise ratio is approximately 2:1.

It should be emphasized again that the role of the forward model $\boldsymbol{f}(\boldsymbol{m})$ is to map (stiffness) parameters $\boldsymbol{m}$ to surface displacement observations $\boldsymbol{d}$. In terms of the forward solution $u(z,t)$, the forward model can be expressed as a

vector with components

$$f_i(\boldsymbol{m}) = u(0, t_i), \qquad i = 1, \ldots, 120, \tag{5.30}$$

where $t_1, \ldots, t_{120}$ are the observation times. Although the underlying PDE is linear, this forward model map from parameters to observables is not.

5.3.2 The likelihood function

The likelihood function governs the probability that a candidate set of stiffness parameters $\mu(z; \boldsymbol{m})$ would reproduce the observation data $\boldsymbol{d}_{\text{obs}}$ that was measured in the exploration experiment. In our case, this observation data is synthetically generated according to the noise model assumed in the previous section.

We generate the experimental observation data on a different mesh than the one used for statistical inversion (256 elements), and we additionally corrupt the observation data with additive Gaussian noise as discussed previously: $\boldsymbol{y}_{\text{obs}} = g(\boldsymbol{\mu}) + \varepsilon_{\text{noise}}$, where $\varepsilon_{\text{noise}} \sim \mathcal{N}(\boldsymbol{0}, \boldsymbol{\Gamma}_{\text{noise}})$, and $\boldsymbol{\Gamma}_{\text{noise}} = \sigma^2_{\text{noise}} \boldsymbol{I}$. This is done to avoid "inverse crimes" [117], in which it might be artificially easy to invert for the desired parameters if the same mesh is used for the inversion as was used to generate the synthetic observation data.

Using the additive Gaussian noise model, our likelihood function is given as

$$\pi_{\text{like}}(\boldsymbol{d}_{\text{obs}} | \boldsymbol{m}) \propto \exp\Big[-\tfrac{1}{2} (\boldsymbol{f}(\boldsymbol{m}) - \boldsymbol{d}_{\text{obs}})^T \boldsymbol{\Gamma}^{-1}_{\text{noise}} (\boldsymbol{f}(\boldsymbol{m}) - \boldsymbol{d}_{\text{obs}}) \Big]. \tag{5.31}$$

For the 2D problem (described in the next subsection), we use $\sigma_{\text{noise}} = 8 \times 10^{-5}$, and for the 16D, 65D, and 1025D problems, we use $\sigma_{\text{noise}} = 2 \times 10^{-5}$.

5.3.3 Parametrizations and priors

In this application, we set up four distinct statistical inverse problems which differ in the parametrization used to describe the medium, and the prior imposed on each parametrization. The choice of prior in a statistical inverse problem can have a significant impact on both the computational effort required to solve the problem as well as the posterior density itself. We select priors in this section that are intended to be typical of the priors that might be used in this case of a heterogeneous medium in which the fine scale variability of the medium is assumed negligible. Toward this end, we use Gaussian smoothness priors, which provide a flexible way to describe random fields with a desired degree of smoothness, and are commonly employed in Bayesian inference of parameter fields. Except in the 1025D case described below, these parametrizations are considered as independent problems, and each have synthetic observation data which are unique to that parametrization.

In the simplest 2D case (i.e. with 2 independent parameters m_1, m_2) the medium is parameterized with four equal length layers, where we constrain the parameter values of the topmost and bottom most layers to be $\mu = 1$, leaving only two degrees of freedom in the parametrization for the second and third layers. In this 2D case, we take the prior to be uniform over $[0.5, 10]$ (i.e. we specify as little a-priori knowledge as possible except for the range of possible

values),

$$\pi_{\text{pr}}(\boldsymbol{m}) \propto \begin{cases} 1 & \text{if} \quad 0.5 \leq m_i \leq 10, \quad \forall i \\ 0 & \text{otherwise} \end{cases} \tag{5.32}$$

In the intermediate 16D case, the medium is parameterized with 16 equal length layers, each containing four elements. In this case we do not further constrain any of the layer parameter values, but we use a (truncated) Gaussian smoothness prior to specify a-priori knowledge that there should not be large jumps between parameter values in neighboring layers. The form of the prior is given explicitly for the layer parameter values m_i for $i = 1, \ldots, 16$.

$$\pi_{\text{prior}}(\boldsymbol{m}) \propto \begin{cases} \exp\Big(-\frac{1}{2}(\boldsymbol{m} - \bar{\boldsymbol{m}}_{\text{prior}})^T \boldsymbol{\Gamma}_{\text{prior}}^{-1} (\boldsymbol{m} - \bar{\boldsymbol{m}}_{\text{prior}}) & 0.5 \leq m_i \leq 10 \quad \forall i \\ 0 & \text{otherwise} \end{cases} \tag{5.33}$$

$$\bar{\boldsymbol{m}}_{\text{prior}}^i = 5 \tag{5.34}$$

$$\Gamma_{\text{prior}}^{ij} = \theta_1 \exp\left(\frac{-(z_i - z_j)^2}{2\theta_2^2}\right) + \varepsilon \delta_{ij} \tag{5.35}$$

The values θ_1 and θ_2 specify the magnitude of the correlation, and the correlation length, respectively. The layer depths z_i indicate the midpoint of each layer. In this example, the correlation length is chosen to be $\theta_2 = 0.125$ (a width of 2 layers). We add a small diagonal term $\varepsilon \delta^{ij}$ to ensure that the prior covariance remains numerically well conditioned. Here we choose $\varepsilon = 10^{-5}$.

Finally, the 65D and 1025D cases allow every discretization point in the mesh to be a separate parameter. (As such, the 1025D problem must be computed on a finer mesh.) We impose the same form of the prior as in the 16D case (again specifying a-priori information about smoothness of the medium), where the z_i represent the depth of each mesh grid point, and we set

the correlation length to $\theta_2 = 0.125$, which is intended to correspond to the same correlation length of 2 layers in the 16D case. As before, we add a small diagonal term $\varepsilon\delta^{ij}$ to ensure numerical well conditioning. We choose $\varepsilon = 10^{-5}$ in the 65D case, and $\varepsilon = 10^{-12}$ for the 1025D case.

5.3.4 The statistical inverse problem

We are now prepared to describe the statistical inverse problem we seek to solve. Sections 5.3.1–5.3.3 describe in detail the ingredients (forward model, likelihood, and prior) required to construct the posterior density using Bayes' theorem,

$$\pi_{\text{post}}(\boldsymbol{m}) \propto \pi_{\text{prior}}(\boldsymbol{m})\pi_{\text{like}}(\boldsymbol{d}_{\text{obs}}|\boldsymbol{m}). \tag{5.36}$$

Complete specification of a particular statistical inverse problem requires a set of observation data $\boldsymbol{d}_{\text{obs}}$ as defined in section 5.3.2, and a choice of medium parametrization and prior as defined in section 5.3.3.

In this paper, we consider three distinct inverse problems, corresponding to different choices of the 2D, 16D, and 65D medium parametrizations and associated prior distributions, as described in section 5.3.3. In each problem, a sample from the prior distribution is selected to be the "ground truth" medium, which is then used to generate synthetic observation data as in section 5.3.2. As an experiment in "weak scaling" of our method, the same observation data are used in the 65D and 1025D experiments. In this sense, the 1025D problem is a refinement of the 65D problem, in which we desire to infer over a larger parameter space for the same fundamental underlying problem.

Finally, "solving" a statistical inverse problem reduces to the ability to interrogate $\pi_{\text{post}}(\boldsymbol{m})$. In high dimensions, this is a nontrivial problem even when the posterior density is known. Typically we are interested in the mean and covariance of the posterior distribution, and higher moments or other functionals of the distribution (e.g., event probabilities) may be desirable as well. Finally, probability distributions for specific quantities of interest (e.g., the softest type of rock in the medium $\min_z \mu(z; \boldsymbol{m})$) are often also essential for decision making purposes once the statistical inverse problem is characterized.

5.3.5 Efficient computation with adjoint methods

For this problem, efficient computation of gradient and Hessian information is crucial. In this section, we give the expressions derived from deterministic PDE constrained optimization, which we use to perform all derivative computations used in the numerical results. For practical reasons, we make little attempt in this paper to justify the expressions given here, but refer the reader to standard references in PDE constrained optimization [23, 112].

Recall that the cost function from deterministic optimization is analogous to the negative log posterior distribution for this problem. We assume the case of a Gaussian prior on the parameters $\boldsymbol{m}$. The negative log posterior is written:

$$-\log \pi_{\text{post}} = \tfrac{1}{2}(\boldsymbol{f}(\boldsymbol{m}) - \boldsymbol{d}_{\text{obs}})^T \boldsymbol{\Gamma}_{\text{noise}}^{-1}(\boldsymbol{f}(\boldsymbol{m}) - \boldsymbol{d}_{\text{obs}}) + \tfrac{1}{2}(\boldsymbol{m} - \bar{\boldsymbol{m}}_{\text{prior}})^T \boldsymbol{\Gamma}_{\text{prior}}^{-1}(\boldsymbol{m} - \bar{\boldsymbol{m}}_{\text{prior}}). \quad (5.37)$$

In the deterministic context, the quantity (5.37) would be minimized as a function of the parameters $\boldsymbol{m}$, subject to the constraint that $u(z,t)$ satisfy the forward model given in the previous section.

The Lagrangian $L(u,p,\boldsymbol{m})$ serves as a tool to solve this constrained minimization problem, where the adjoint solution $p(z,t)$ is introduced as a Lagrange multiplier to enforce the given constraints.

To write the Lagrangian for the constrained optimization problem, we introduce the adjoint solution $p(z,t)$ which plays the role of the Lagrange multiplier. The Lagrangian can now be expressed in weak form in terms of the forward solution u, the adjoint solution p, and the parameters $\boldsymbol{m}$:

$$
\begin{aligned}
L(u,p,\boldsymbol{m}) = &\int_{t=0}^{t=T}\int_{z\in\Omega}\frac{1}{2\sigma^2_{\text{noise}}}\sum_{i=1}^{120}(u(z,t)-\boldsymbol{d}^i_{\text{obs}})^2\delta(z-0)\delta(t-t_i)\\
&+\frac{1}{2}(\boldsymbol{m}-\bar{\boldsymbol{m}}_{\text{prior}})^T\boldsymbol{\Gamma}^{-1}_{\text{prior}}(\boldsymbol{m}-\bar{\boldsymbol{m}}_{\text{prior}})\mathrm{d}z\mathrm{d}t\\
&+\int_{t=0}^{t=T}\int_{z\in\Omega}p(z,t)\left[\rho u_{tt}(z,t)-\Big(\mu(z;\boldsymbol{m})\,u_z(z,t)\Big)_z-\delta(z-0)F^{\text{ricker}}(t)\right]\mathrm{d}z\mathrm{d}t\\
&+\int_{t=0}^{t=T}\left[p(L,t)\left(\sqrt{\rho\,\mu(L;\boldsymbol{m})}\,u_t(L,t)+\mu(L;\boldsymbol{m})\,u_z(L,t)\right)-p(0,t)\mu(0;\boldsymbol{m})u_z(0,t)\right]\mathrm{d}t\\
&+\int_{z\in\Omega}\rho\left[p(z,0)u_t(z,0)-p_t(z,0)u(z,0)\right]\mathrm{d}z\\
&+\sqrt{\rho\mu(L;\boldsymbol{m})}\,p(L,0)u(L,0).
\end{aligned}
$$

Setting $\delta_p L(u,p,\boldsymbol{m})=0$, we recover the original forward PDE with the proper boundary conditions, given in section 5.3.1.

Setting $\delta_u L(u,p,\boldsymbol{m})=0$ and carefully integrating by parts, we derive the adjoint PDE and boundary conditions which determine the adjoint solution

$p(z,t)$:

$$\rho\, p_{tt}(z,t) - \Big(\mu(z;\boldsymbol{m})\, p_z(z,t)\Big)_z = -\frac{1}{\sigma^2_{\text{noise}}} \sum_{i=1}^{120} (u(z,t) - \boldsymbol{d}^i_{\text{obs}})\delta(z-0)\delta(t-t_i) \qquad \text{(Adj. PDE)}$$

$$\mu(L;\boldsymbol{m})\, p_z(L,t) = \sqrt{\rho\, \mu(L;\boldsymbol{m})}\, p_t(L,t) \qquad \text{(Absorbing BC)}$$

$$\mu(0;\boldsymbol{m})\, p_z(0,t) = 0 \qquad \text{(Free Surface BC)}$$

$$p(z,T) = 0 \qquad \text{(FC)}$$

$$p_t(z,T) = 0 \qquad \text{(FC)}$$

The gradient $\boldsymbol{g} = \nabla_{\boldsymbol{m}} L(u,p,\boldsymbol{m})$ is then computed efficiently using the forward $u(z,t)$ and adjoint $p(z,t)$ functions satisfying the forward and adjoint equations respectively:

$$\begin{aligned} \boldsymbol{g} = \boldsymbol{\Gamma}^{-1}_{\text{prior}}(\boldsymbol{m} - \bar{\boldsymbol{m}}_{\text{prior}}) \\ + \int_0^T \int_\Omega \Big[\nabla_{\boldsymbol{m}} \mu(z;\boldsymbol{m})\Big] p_z(z,t) u_z(z,t)\, \mathrm{d}z\mathrm{d}t \\ + \int_0^T \frac{1}{2}\sqrt{\frac{\rho}{\mu(L;\boldsymbol{m})}}\, \Big[\nabla_{\boldsymbol{m}} \mu(L;\boldsymbol{m})\Big] p(L,t) u_t(L,t) \mathrm{d}t \end{aligned} \tag{5.38}$$

We next consider the block form of the full Hessian operator $\boldsymbol{H} = \nabla^2 L(u,p,\boldsymbol{m})$, which acts on the *incremental variables* $(\tilde{p},\tilde{u},\tilde{\boldsymbol{m}})$. Since only the parameters $\tilde{\boldsymbol{m}}$ are of interest, we consider the Schur complement of $\boldsymbol{H}$ with respect to the $\tilde{\boldsymbol{m}}$ block, which amounts to block elimination of the system

$$\begin{pmatrix} H_{pp} & H_{pu} & H_{p\boldsymbol{m}} \\ H_{up} & H_{uu} & H_{u\boldsymbol{m}} \\ H_{\boldsymbol{m}p} & H_{\boldsymbol{m}u} & H_{\boldsymbol{m}\boldsymbol{m}} \end{pmatrix} \begin{pmatrix} \tilde{p} \\ \tilde{u} \\ \tilde{\boldsymbol{m}} \end{pmatrix} = \begin{pmatrix} 0 \\ 0 \\ \nabla_{\boldsymbol{m}} \nabla_{\boldsymbol{m}} L(u,p,\boldsymbol{m})\tilde{\boldsymbol{m}} \end{pmatrix}, \tag{5.39}$$

which implicitly defines the action of the Hessian on $\tilde{\boldsymbol{m}}$. Again, the details for the justification of the above expression are omited, and we refer the reader to standard references in PDE constrained optimization.

To compute the action of the reduced Hessian operator on a given parameter function $\tilde{\boldsymbol{m}}(z)$, we first solve the *incremental forward equation*, given by row 1 of (5.39):

$$\rho\,\tilde{u}_{tt}(z,t) - \Big(\mu(z;\boldsymbol{m})\,\tilde{u}_z(z,t)\Big)_z = \Big(\Big[\nabla_{\boldsymbol{m}}\mu(z;\boldsymbol{m})\cdot\tilde{\boldsymbol{m}}\Big]\,u_z(z,t)\Big)_z \qquad \text{(Incremental PDE)}$$

$$\mu(L;\boldsymbol{m})\tilde{u}_z(L,t) + \Big[\nabla_{\boldsymbol{m}}\mu(L;\boldsymbol{m})\cdot\tilde{\boldsymbol{m}}\Big]\,u_z(L,t) = \qquad \text{(Bottom BC)}$$
$$-\sqrt{\rho\,\mu(L;\boldsymbol{m})}\,\tilde{u}_t(L,t) - \frac{1}{2}\sqrt{\frac{\rho}{\mu(L;\boldsymbol{m})}}\Big[\nabla_{\boldsymbol{m}}\mu(L;\boldsymbol{m})\cdot\tilde{\boldsymbol{m}}\Big]u_t(L,t)$$

$$\mu(0;\boldsymbol{m})\,\tilde{u}_z(0,t) + \Big[\nabla_{\boldsymbol{m}}\mu(0;\boldsymbol{m})\cdot\tilde{\boldsymbol{m}}\Big]u_z(0,t) = 0 \qquad \text{(Top BC)}$$

$$\tilde{u}(z,0) = 0 \qquad \text{(IC)}$$

$$\tilde{u}_t(z,0) = 0 \qquad \text{(IC)}$$

Next we solve the *incremental adjoint equation*, given by row 2 of (5.39):

$$\begin{aligned}
\rho\,\tilde{p}_{tt}(z,t) - \Big(\mu(z;\boldsymbol{m})\,\tilde{p}_z(z,t)\Big)_z = \quad & \text{(Incremental Adj. PDE)} \\
\Big(\Big[\nabla_{\boldsymbol{m}}\mu(z;\boldsymbol{m})\cdot\tilde{\boldsymbol{m}}\Big]\,p_z(z,t)\Big)_z - \frac{1}{\sigma^2_{\text{noise}}}\sum_{i=1}^{120}\tilde{u}\delta(z-0)\delta(t-t_i) &
\end{aligned}$$

$$\begin{aligned}
\mu(L;\boldsymbol{m})\tilde{p}_z(L,t) + \Big[\nabla_{\boldsymbol{m}}\mu(L;\boldsymbol{m})\cdot\tilde{\boldsymbol{m}}\Big]p_z(L,t) = \quad & \text{(Bottom BC)} \\
\sqrt{\rho\,\mu(L;\boldsymbol{m})}\tilde{p}_t(L,t) + \frac{1}{2}\sqrt{\frac{\rho}{\mu(L;\boldsymbol{m})}}\Big[\nabla_{\boldsymbol{m}}\mu(L;\boldsymbol{m})\cdot\tilde{\boldsymbol{m}}\Big]p_t(L,t) & \\
\mu(0;\boldsymbol{m})\,\tilde{p}_z(0,t) + \Big[\nabla_{\boldsymbol{m}}\mu(0;\boldsymbol{m})\cdot\tilde{\boldsymbol{m}}\Big]p_z(0,t) = 0 \quad & \text{(Top BC)} \\
\tilde{p}(z,T) = 0 \quad & \text{(FC)} \\
\tilde{p}_t(z,T) = 0 \quad & \text{(FC)}
\end{aligned}$$

Finally, the *Hessian-vector product* can be computed by row 3 of (5.39), using the solutions of the incremental forward and adjoint equations, respectively:

$$\begin{aligned}
\boldsymbol{H}\tilde{\boldsymbol{m}} = \boldsymbol{\Gamma}^{-1}_{\text{prior}}\tilde{\boldsymbol{m}} & \\
+ \int_0^T\int_\Omega \Big\{\Big[\nabla_{\boldsymbol{m}}\mu(z;\boldsymbol{m})\Big]\Big(\tilde{p}_z(z,t)u_z(z,t) + p_z(z,t)\tilde{u}_z(z,t)\Big) & \\
+ \Big[\nabla^2_{\boldsymbol{m}}\mu(z;\boldsymbol{m})\cdot\tilde{\boldsymbol{m}}\Big]p_z(z,t)u_z(z,t)\Big\}\mathrm{d}z\mathrm{d}t & \\
+ \int_0^T\Big\{\frac{1}{2}\sqrt{\frac{\rho}{\mu(L;\boldsymbol{m})}}\Big[\nabla_{\boldsymbol{m}}\mu(L;\boldsymbol{m})\Big]\Big(\tilde{p}(L,t)u_t(L,t) + p(L,t)\tilde{u}_t(L,t)\Big) & \\
- \frac{1}{4}\sqrt{\frac{\rho}{\mu(L;\boldsymbol{m})^3}}\Big[\nabla^T_{\boldsymbol{m}}\mu(L;\boldsymbol{m})\Big]\Big[\nabla_{\boldsymbol{m}}\mu(L;\boldsymbol{m})\cdot\tilde{\boldsymbol{m}}\Big]p(L,t)u_t(L,t) & \\
+ \frac{1}{2}\sqrt{\frac{\rho}{\mu(L;\boldsymbol{m})}}\Big[\nabla^2_{\boldsymbol{m}}\mu(L;\boldsymbol{m})\cdot\tilde{\boldsymbol{m}}\Big]p(L,t)u_t(L,t)\Big\}\mathrm{d}t &
\end{aligned}$$

5.4 Numerical results

The primary goal of this section is to compare the performance of a variety of MCMC methods for the statistical inverse problems outlined in Section 5.3.4. Four separate statistical inverse problems are considered, corresponding to the choices of of the medium parametrizations and associated priors outlined in Section 5.3.3. We call these experiments 2D, 16D, 65D, and 1025D, respectively, according to the number of parameter dimensions contained in the problem.

The observation data $\boldsymbol{d}_{\text{obs}}$ for each of the experiments are synthetically generated using a "ground truth" medium in each case which is drawn from the prior distribution. Synthetic observation data is generated on a different mesh than the one used for inversion.

In general the datasets used in each case are unique, with the specific exception that the 1025D problem is intended to be a precise refinement of the 65D problem: The same dataset is used in both experiments, and the initial starting points for the MCMC chains in the 1025D case are linear interpolations of the starting points for the 65D case. In this way, the same underlying *physical* problem is being solved in both cases, so that we may look at the scaling behavior of our method.

For each MCMC method to be compared, 64 MCMC chains are computed using a common set of 64 initial points. These points are selected from a long Stochastic Newton MCMC chain which is initialized at the MAP esti-

mate. From this chain, several initial points are chosen which approximately maximize the minimum pairwise distances between points, so that the resulting set is distributed quasi-uniformly over the region of non-negligible posterior probability density. In this way the initial points are over-dispersed relative to the true posterior probability density (Which is important for computation of the MPSRF, later), but removes potential difficulties in comparing different "burn-in" times for different MCMC methods, and makes the results more comparable in general. In cases where the MCMC method requires tuning or choice of parameters, several parameter studies were performed to attempt to optimize the performance of the MCMC chain, wherein we choose the parameter(s) which provide the largest mean square jump distance while maintaining an acceptance rate of 30%-50% [177].

Secondary goals are to demonstrate features of this particular physical model which enable the use of Stochastic Newton MCMC, and to examine quantities of interest which might be of scientific or engineering relevance.

5.4.1 Visualization of the posterior pdf

Attempting to construct a visualization which depicts the full correlation structure for a 65-dimensional object is an impossible task. In this section, we present the most generally informative visualization of the solution to the given statistical inverse problems that we are able to provide.

In Figure 5.2, we present the marginalized one-dimensional probability distributions as a vertical gray scale stripe for each depth. Regions of darker

gray indicate higher certainty that the true curve passes through a given value of the parameter at this depth. The images in the figure are constructed by placing these gray scale stripes side by side for every depth, and as such present no indication of the correlation between parameter values at different depths.

To give a hint at the correlation structure, a few representative samples are shown, drawn from the prior PDF, or drawn from the posterior PDF MCMC chain, respectively. In all cases, the blue curve represents the ground truth parameters, from which the synthetic observations were generated, and should be expected to pass through the regions of reasonable (or at least non-negligible) probability.

5.4.2 MPSRF diagnostic

To compare the different MCMC methods, we employ the multivariate potential scale reduction factor diagnostic (MPSRF) [24]. This diagnostic compares averaged properties of the individual sample chains with properties of the pooled sample chain. When these properties are similar, we infer that each of the individual sample chains has converged.

This idea is made quantitative using the sample chain covariance. One estimate $\boldsymbol{W}$ uses the average of the individual sample chain covariances, which will tend to underestimate the true covariance of the distribution. Second, $\hat{\boldsymbol{V}}$ estimates the pooled sample chain covariance between all of the chains, and will tend to overestimate the true sample covariance, due to the over-dispersion of the initial points. The MPSRF statistic then computes the maximum linear

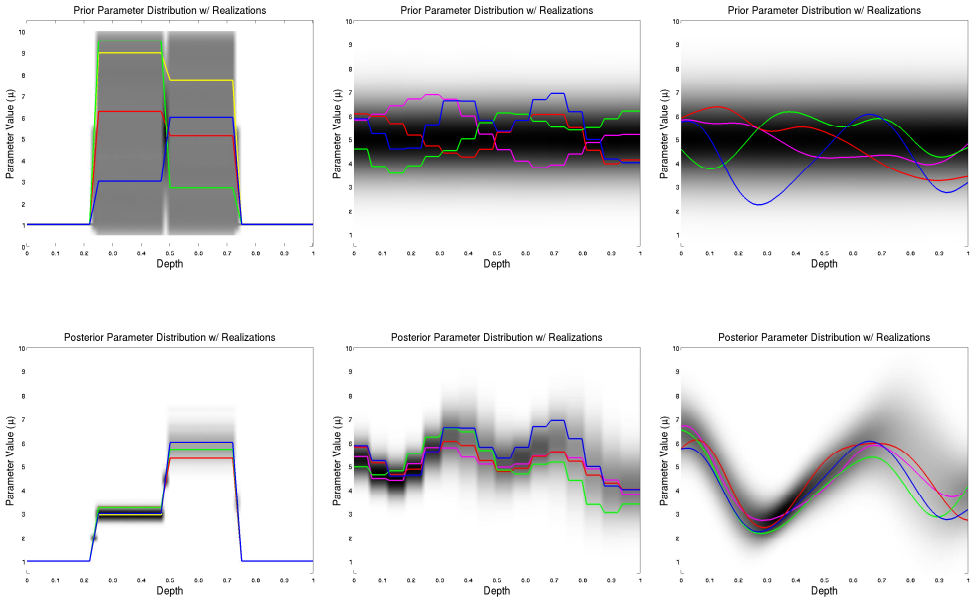

Figure 5.2: Visualizations of the 1-D marginal prior probability distributions (top row) and posterior probability distributions (bottom row) are shown in gray scale above. Results are shown from left to right for each of the 2, 16, and 65-dimensional parametrizations of the medium, respectively. A few realizations from each distribution are overlaid to give indication of the smoothness of the distributions. Parametrizations shown in blue on each of the plots represent the "true" underlying distribution from which the observation data were generated.

projection of the ratio,

$$\sqrt{\hat{R}} = \max_{\|\boldsymbol{a}\|=1} \left(\frac{\boldsymbol{a}^T \hat{\boldsymbol{V}} \boldsymbol{a}}{\boldsymbol{a}^T \boldsymbol{W} \boldsymbol{a}} \right), \tag{5.40}$$

which overestimates and eventually approaches 1 as the two estimates become more and more similar.

Figure 5.3 displays the MPSRF curves for each MCMC method on each of the 2D, 16D, and 65D problems. We find in general that Stochastic Newton MCMC is always more efficient on a per sample basis, but the reference MCMC methods DRAM and Langevin are very competitive on the 2D and 16D problems in terms of total computation time, as Stochastic Newton is substantially more expensive. However, the reference methods fail to converge for the 65D problem under the MPSRF convergence diagnostic even in 10 hours of wallclock computation time and order 10^5 samples, while Stochastic Newton does still appear to converge.

In Figure 5.4, we demonstrate scaling of the low rank Stochastic Newton to large-scale problems, by comparing the MPSRF convergence diagnostic for the 65D and 1025D problems plotted against the number of samples computed. If the 65D problem is well resolved, then we anticipate that the 1025D should display similar convergence diagnostics (as a function of number of samples), since it is in principle nothing more than a refinement of the same problem. Furthermore, we have claimed previously that each sample requires a dimension independent number of PDE solves (depending only on the compact subspace of the Hessian), and therefore the full solution cost for the statisti-

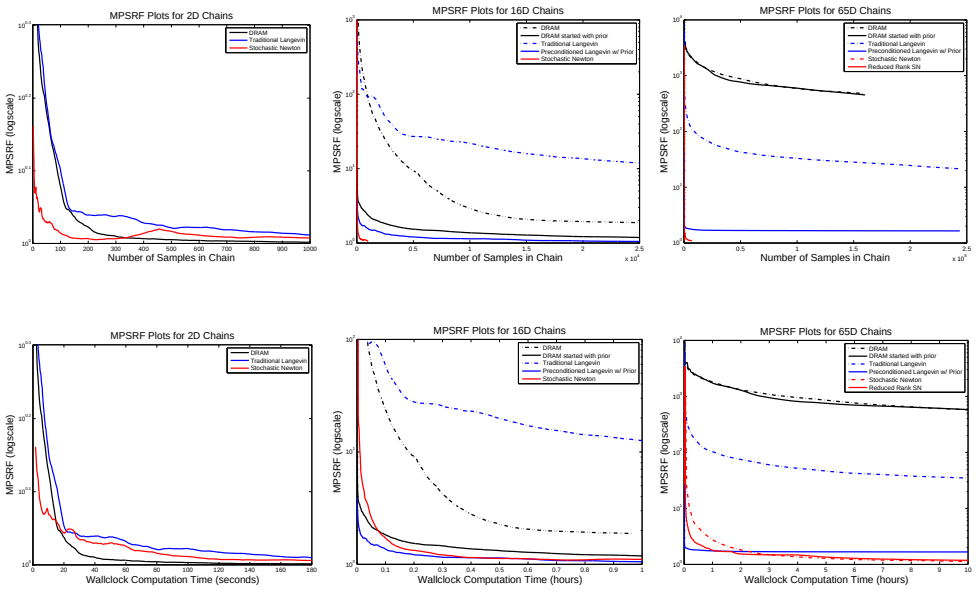

Figure 5.3: The MPSRF statistic is shown on a semi-logarithmic plot for Stochastic Newton MCMC (in red), and two reference MCMC methods DRAM (in black) and Langevin (in blue). As convergence is reached, the MPSRF estimate is expected to decrease to 1. The MPSRF is plotted as a function of the number of samples in each of 64 parallel MCMC chains (top row), and as a function of the total per-chain wallclock computation time (bottom row). Stochastic Newton generally requires several PDE solves for each MCMC sample while the reference methods DRAM and Langevin only require 1 and 2 PDE solves, respectively, which accounts precisely for the differences in the top and bottom rows. Results are shown from left to right for each of the 2-, 16-, and 65-dimensional parametrizations of the medium, respectively. In the smaller problems (left and middle columns), the reference MCMC methods are very competitive with Stochastic Newton. However in the largest problem (right column), the reference MCMC methods fail to even converge in 10 hours of wallclock computation time and $O(10^5)$ samples under this metric.

cal inverse problem should be only a constant multiple of the cost of a single forward PDE solve, which is independent of the parameter dimension.

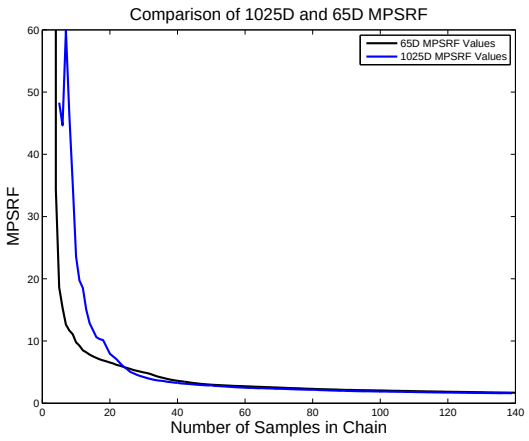

Figure 5.4: The MPSRF statistic for the 1025D and 65D Stochastic Newton MCMC chains is plotted as a function of the number of samples in each chain. We observe similar convergence rates for both problems in this metric, despite the factor of 16 difference in number of parameters. This demonstrates that Stochastic Newton is only sensing the intrinsic difficulty of the problem defined by the compact subspace of the Hessian operator, rather than the full 65 or 1025 parameter dimensions, which are otherwise fatal for the reference methods.

5.4.3 MCMC chain statistics

Table 5.1 shows the computational time per sample, mean square jump distance, and integrated autocorrelation times for six scalar quantities of interest.

Time per sample Comparing the computational time required for eah MCMC sample (TPS column of Table 5.1), we see that full Stochastic Newton

			Integrated Autocorrelation Times					
MCMC type	TPS	MSJ	$\min \mu(z)$	$\max \mu(z)$	$\int_0^L \mu(z)\mathrm{d}z$	$\mu(0)$	$\mu(L/2)$	$\mu(L)$
SN	64	6.1	65	124	50	17	52	31
rr SN	16	6.8	85	95	46	37	56	32
L	0.42	3.0e-4	–	–	–	–	–	–
pp L	0.42	5.9	74	114	52	29	51	35
DRAM	0.35	1.2e-5	–	–	–	–	–	–
pi DRAM	0.35	1.2e-5	–	–	–	–	–	–

Table 5.1: Time per sample in seconds (TPS), mean squared jump distance (MSJ) and integrated autocorrelation time comparison for a variety of MCMC methods. We compare the full rank Stochastic Newton MCMC (SN), reduced rank Stochastic Newton (rr SN), Langevin MCMC (L), prior-preconditioned Langevin MCMC (pp L), Delayed Rejection Adaptive Metropolis MCMC (DRAM), and prior-initialized DRAM (pi DRAM). Entries for which integrated autocorrelation is not listed are incomputable due to lack of chain convergence. The 65 parameter experiment is considered for all statistics.

easily has the highest per sample expense followed by reduced rank Stochastic Newton. The reference MCMC methods are comparatively inexpensive.

Mean square jump distance The mean square jump distance (MSJ column of Table 5.1) can also be used to give an indication of how well the MCMC chain is mixing within the desired posterior probability distribution. This metric is defined for a single MCMC chain with samples $\boldsymbol{m}_0, \ldots, \boldsymbol{m}_N$ as

$$\text{MSJ} := \frac{1}{N} \sum_{k=0}^{N-1} \|\boldsymbol{m}_{k+1} - \boldsymbol{m}_k\|^2. \tag{5.41}$$

The quantity reported in the table is averaged among all 64 parallel chains from a given method. In general, a larger mean square jump distance indicates faster mixing of the MCMC chain, and tends to result in better chain convergence

to the underlying posterior distribution.

Integrated autocorrelation time In the 65D case, we also consider the integrated autocorrelation times for six scalar quantities of interest for each MCMC method considered. These quantities are computed for each sample as $\min_z \mu(z;\boldsymbol{m})$, $\max_z \mu(z;\boldsymbol{m})$, $\int_0^L \mu(z;\boldsymbol{m})\mathrm{d}z$, $\mu(0;\boldsymbol{m}), \mu(L/2;\boldsymbol{m})$, and $\mu(L;\boldsymbol{m})$, respectively, and reported in Table 5.1.

It is well known in Monte Carlo methods that averaging over N i.i.d. samples $\boldsymbol{m}_k \sim \pi_{\text{post}}$ will reduce the variance in the estimate by a factor of $\frac{1}{N}$. However, MCMC samples are most certainly not independent, and in general we observe that averaging over N samples from an MCMC chain will reduce the variance in the estimate by a factor of only $\frac{\tau}{N}$, where $\tau > 1$ is the integrated autocorrelation time [174]. This can be computed as

$$\tau = 1 + 2\sum_{s=1}^{\infty} \rho(s), \tag{5.42}$$

where $\rho(s)$ is the usual autocorrelation function for a lag s. In practice for finite length sample chains, $\rho(s)$ is a noisy function, and we report the maximum value of τ obtained by truncating the summation after any value of $s < 5000$. In some cases, the sum does not converge over the entire length of the sample chain. It is almost certain that these chains are not well converged, and these entries not reported.

We observe similar integrated autocorrelation times for the full and reduced rank Stochastic Newton MCMC methods, as well as the prior preconditioned Langevin MCMC, indicating that these methods appear to be mixing

well and at comparable rates. In this metric we actually do not observe significant difference between Stochastic Newton and prior preconditioned Langevin MCMC, which is likely due to the smoothing effects in our choice of prior.

5.4.4 Compactness of the likelihood Hessian

We demonstrate numerically that the Hessian matrix of the likelihood term for this problem is indeed compact, as this is a necessary condition for Stochastic Newton MCMC to be effective. Figure 5.5 contains spectra for the likelihood Hessian $\boldsymbol{H}_{\text{misfit}} = -\nabla^2 \log \pi_{\text{like}}$ computed at each of the 64 MCMC chain starting points.

The spectra shown do not contain the contribution to the Hessian from the prior term and are not modified (e.g., for positive definiteness) to demonstrate the low rank character of the underlying physical model.

It is precisely this underlying compact nature of the forward model that enables similar convergence characteristics of the refined 1025D problem as those of the 65D problem, demonstrated in Figure 5.4.

5.5 Concluding remarks

We have presented a Stochastic Newton MCMC method that is aimed at ill-posed and large-scale statistical inverse problems. The key idea is to make use of gradient and Hessian information characterizing the posterior probability density function. We apply concepts from deterministic optimization, making the connection to the classical Newton's method, to efficiently

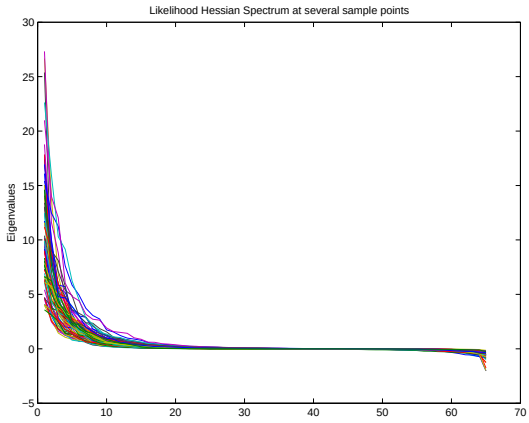

Figure 5.5: Likelihood Hessian spectrum curves, computed at each of 64 sample points distributed quasi-uniformly across the region of non-negligible posterior pdf. The spectrum at every point collapses quickly to zero, as even a single noncompact spectrum would stand out among the rest in this view. We also observe small negative eigenvaleus in some spectra, demonstrating nonlinearity and nonconvexity of our forward model.

construct a proposal density for MCMC sampling without ever building the full Hessian operators.

We apply the proposed method to a prototypical statistical inverse problem based on a seismological scattering experiment that is governed by a 1D wave equation. Stochastic Newton MCMC and two reference MCMC methods are applied to this problem for a variety of discretizations of the parameter space. When the number of parameters is small, all three methods are comparable in MCMC performance. However, when increasing the dimension of the parameter space, Stochastic Newton shows faster convergence and better mixing of the MCMC chain. Moreover, comparing its performance for 65D and 1025D parametrizations of the same physical problem, we observe

similar MCMC convergence characteristics. While this behavior is not yet provable theoretically, the numerical observations suggest an insensitivity of convergence of Stochastic Newton to the parameter dimension.

We hypothesize that the observed dimension-independence of the proposed method (depicted in Figure 5.4) stems from its ability to detect the subspace of parameters for which the data are informative (and therefore the forward model is active), which is typically small for ill-posed inverse problems governed by PDEs. Once this data-informed subspace is sufficiently well resolved by a given parameter discretization, we anticipate that further parameter refinement does not affect the data misfit term, and therefore does not affect the resulting posterior distribution or the low-rank character of the Hessian. We thus expect the number of PDE solves required for Stochastic Newton MCMC to be similarly unaffected as the parameter dimension is increased, enabling this method to be effective for PDE-based statistical inverse problems with high-dimensional parameter spaces.

We are currently investigating robustness of the method with respect to nonlinearity of the forward model and indefiniteness of the Hessian, which both can produce high sample rejection rates.

Acknowledgments

The authors would like to thank Youssef Marzouk, Tan Bui-Thanh, Georg Stadler, Ernesto Prudencio, Todd Oliver, Karl Schulz, and Colin Fox for their many engaging discussions and helpful advice over the course of this

work. We thank the referees for their careful reading of the manuscript and their helpful suggestions. The most intensive calculations were performed on the Ranger supercomputer at the Texas Advanced Computing Center under NSF TeraGrid award TG-MCA04N026. This research was supported by AFOSR grant FA9550-09-1-0608, NSF grants DMS-0724746, ARC-0941678, and CMMI-1028889, and DOE grants DE-FG02-08ER25860, DE-SC0002710, and DE-FC52-08NA28615. James Martin was supported by the DOE CSGF under grant number DE-FG02-97ER25308.

Chapter 6

A computational framework for infinite-dimensional Bayesian inverse problems Part II: Stochastic Newton MCMC with application to ice sheet flow inverse problems

The content of this chapter is based on an existing publication[1] which is joint work with Noemi Petra, Georg Stadler, and my advisor Omar Ghattas. Noemi and Georg both made significant contributions to the development of a framework for solving inverse problems in COMSOL. Noemi implemented the particular forward and adjoint codes for the Arolla test problem. Noemi and I worked together on the implementation of the statistical algorithms. Noemi also contributed the setup and running of each batch of numerical experiments, followed by collaborative efforts with Georg and myself to interrogate the results, assess convergence, visualize the data, and select the next experiments to be run. Most of the writing for the introduction and description of the Arolla test problem was done by my co-authors. All authors had significant contribution to the remaining content of this chapter.

[1] N. Petra, J. Martin, G. Stadler, and O. Ghattas. A computational framework for infinite-dimensional Bayesian inverse problems: Part II. Stochastic Newton MCMC with application to ice sheet inverse problems. *SIAM Journal on Scientific Computing*, 36(4):A1525–A1555, 2014. http://epubs.siam.org/doi/abs/10.1137/130934805

Abstract

We address the numerical solution of infinite-dimensional inverse problems in the framework of Bayesian inference. In the Part I [33] companion to this paper, we considered the linearized infinite-dimensional inverse problem. Here in Part II, we relax the linearization assumption and consider the fully nonlinear infinite-dimensional inverse problem using a Markov chain Monte Carlo (MCMC) sampling method. To address the challenges of sampling high-dimensional probability density functions (pdfs) arising upon discretization of Bayesian inverse problems governed by PDEs, we build on the stochastic Newton MCMC method. This method exploits problem structure by taking as a proposal density a local Gaussian approximation of the posterior pdf, whose covariance operator is given by the inverse of the local Hessian of the negative log posterior pdf. The construction of the covariance is made tractable by invoking a low-rank approximation of the data misfit component of the Hessian. Here we introduce an approximation of the stochastic Newton proposal in which we compute the low-rank-based Hessian at just the MAP point, and then reuse this Hessian at each MCMC step. We compare the performance of the proposed method to the original stochastic Newton MCMC method and to an independence sampler. The comparison of the three methods is conducted on a synthetic ice sheet inverse problem. For this problem, the stochastic Newton MCMC method with a MAP-based Hessian converges at least as rapidly as the original stochastic Newton MCMC method, but is far cheaper since it avoids recomputing the Hessian at each step. On the other hand, it is more

expensive per sample than the independence sampler; however, its convergence is significantly more rapid, and thus overall it is much cheaper. Finally, we present extensive analysis and interpretation of the posterior distribution, and classify directions in parameter space based on the extent to which they are informed by the prior or the observations.

6.1 Introduction and background

We consider the problem of estimating the uncertainty in the solution of infinite-dimensional inverse problems within the framework of Bayesian inference [117, 189, 192]. Namely, given observational data and their uncertainties, a (possibly stochastic) forward model that maps model parameters to observations, and a prior probability distribution on model parameters that encodes any prior knowledge or assumptions about the parameters, find the *posterior probability distribution* of the parameters conditioned on the observational data. This probability density function (pdf) is defined as the Bayesian solution of the inverse problem. The posterior distribution assigns to any candidate set of parameter fields our belief (expressed as a probability) that a member of this candidate set is the "true" parameter field that gave rise to the observed data.

The standard approach to explore the posterior distribution is based on sampling using a Markov chain Monte Carlo (MCMC) method. However, the use of conventional MCMC methods becomes intractable for large-scale inverse problems, which arise upon discretization of infinite-dimensional in-

verse problems. This is due to the twin difficulties of high dimensionality of the uncertain parameters and computationally expensive forward models.

A number of methods have emerged to address Bayesian inverse problems governed by PDEs (we give a representative recent reference in each case, which can be consulted for additional references to historical work; further references can be found in the recent survey [76]): replacing the forward problem with a reduced order model in both parameter and state space [136]; approximating the parameter-to-observable map [109] or the posterior [32] with a Gaussian process response surface; employing a polynomial chaos approximation of the forward problem [145]; using a two-stage "delayed acceptance" MCMC method in which the first stage employs an approximate forward model [55]; employing gradient information (of the negative log posterior) to accelerate sampling, as in Langevin methods [61, 180, 190] and their preconditioned variants [19]; exploiting Riemannian geometry of parameter space to accelerate sampling [86]; and creating an MCMC proposal that uses local gradient and low-rank Hessian information of the negative log posterior to construct a local Gaussian approximation [143].

Here we focus on the last of these methods, the so-called *stochastic Newton MCMC method*. This method employs local Hessian-based Gaussian proposals that exploit the structure of the underlying posterior to guide the sampler to regions with higher acceptance probability. In particular, such proposals capture the highly stretched contours of the posterior that are typical for ill-posed inverse problems, in which the data inform the model parame-

ters very well in some directions in parameter space, and poorly in others. One of the challenges in employing the Hessian is that its explicit construction entails solution of as many forward problems as there are parameters, which is out of the question for large-scale forward problems. These difficulties are addressed by introducing low-rank approximations of the Hessian, motivated by the compact nature of the Hessian operator for many inverse problems [26, 29, 30, 31, 33, 73, 143]. This delivers accurate approximation of the Hessian at a cost that is independent of the parameter dimension (when the parameter represents a discretized field), leading to solution of Bayesian inverse problems with non-trivial dimensions [143]. Other work employing Hessian-based proposals includes the taylored chain approach [82], and, specifically, in the context of nonlinear filtering [81]; the Hessian-based Metropolis-Hastings (HMH) algorithm with a learning rate to influence step size [171]; a position-specific preconditioned Metropolis adjusted Langevin algorithm (PSP-MALA) implemented with a block Metropolis-Hastings algorithm [106]; function-space MCMC proposals for which the prior is invariant, and thus insensitive to mesh refinement [126]; and, finally, the Random Maximized Likelihood (RML) algorithm which generates samples as the solutions of related deterministic inverse problems [157].

Despite the low-rank approximation, stochastic Newton MCMC (and any method that uses local Hessian information) is computationally expensive for large-scale problems, since at every proposed sample point the gradient and a low-rank approximation of the Hessian are computed, which requires

multiple forward and adjoint PDE solves having the same linear operator (or its adjoint). When these PDE solves are done iteratively, there is little opportunity to exploit the fact that the linear operators are the same, beyond amortizing the cost of preconditioner construction over the solves.

To alleviate this computational cost, here we propose a modified stochastic Newton MCMC that uses proposals based on local gradient information as well as on Hessian information computed initially at the maximum a posteriori (MAP) point and then reused at every sample point. We call this the *stochastic Newton MCMC method with MAP-based Hessian*. We compare this proposed method with the original stochastic Newton MCMC method (with dynamically-computed Hessian) as well as with an independence sampler that uses a Gaussian proposal centered at the MAP point, using the Hessian computed at the MAP as the covariance [157]. This independence sampler is computationally attractive since (like the proposed stochastic Newton MCMC method with MAP-based Hessian), the Hessian is computed just once, but (unlike the new method), the gradient is used only to determine the MAP. Because the proposed stochastic Newton MCMC method with MAP-based Hessian uses local (gradient) information, we expect it will outperform the independence sampler; because it freezes the Hessian at the MAP point, it will be significantly cheaper per sample than the original stochastic Newton MCMC method.

The stochastic Newton MCMC method with MAP-based Hessian can be derived as a particular variant of a preconditioned Metropolis-adjusted

Langevin algorithm using preconditioning based on the Hessian at the MAP point. Note that all of the above methods attempt to exploit problem structure—in particular the local curvature of the posterior—by making use of Hessian information to one degree or another. Note also that all three of these Hessian-based methods reduce to the same method when the target inverse problem is linear and the prior and noise pdfs are Gaussian (in which case the posterior is also Gaussian). For non-Gaussian posteriors, however, the three methods take distinct steps.

Beyond this new, more efficient, variant of stochastic Newton MCMC, this article extends our previous work on methods for large-scale Bayesian inverse problems [33, 143] in several directions. In [33], we presented a computational framework for linearized infinite-dimensional Bayesian inverse problems, building on the infinite-dimensional formulation of Stuart [189]. Here, we extend our computational framework to nonlinear inverse problems, for which the posteriors are non-Gaussian, requiring MCMC sampling. To this end, we extend the finite-dimensional stochastic Newton MCMC method presented in [143] to be consistent with the infinite-dimensional setting. This requires care in discretizing the prior and likelihood and establishing finite-dimensional inner products, which arise in multiple steps of stochastic Newton.

We study the efficiency of the proposed method in the context of an ice sheet flow Bayesian inverse problem, in which a basal boundary condition parameter field is inferred from surface velocity observations. Here, the parameter-to-observable map involves the solution of a nonlinear Stokes equa-

tion describing viscous, creeping, incompressible, non-Newtonian ice flow. This extends recent research on ice sheet inverse problems, which focused on deterministic inversion or the computation of the MAP solution [88, 151, 167, 169, 170, 172]. We apply the full Bayesian inference framework and study the performance of the three Hessian-based methods described above in exploring the posterior pdf. Convergence of the three methods is studied using various diagnostics to assess MCMC chain convergence. We also compare with a reference Delayed Rejection Adaptive Metropolis (DRAM) sampler [95] that, similar to stochastic Newton, attempts to capture the curvature of the posterior, but without relying on gradient or Hessian information. The results reveal that, among the Hessian-based methods, the stochastic Newton MCMC method with MAP-based Hessian yields the fastest convergence in terms of both the number of samples and the computational work. In comparison, DRAM is incapable of making progress on this problem.

Finally, we study and interpret visually the solution of the Bayesian inverse problem with respect to the information contained in the data and in the prior and the effect they have on the posterior in high dimensions. This can be challenging in high dimensions, but we demonstrate that it can be made tractable by exploiting knowledge contained within the spectral structure of the Hessian of the log likelihood evaluated at the MAP point as well as the prior covariance. Because this structure is common to many Bayesian inverse problems, we expect that these strategies for visualization will be of general value beyond the specific application.

The remaining sections of this paper are organized as follows. We begin by providing in Section 6.2.1 an overview of the framework for infinite-dimensional Bayesian inverse problems following [33,189]. Next, in Section 6.2.2 we present a consistent discretization of the infinite-dimensional inverse problem. Section 6.3 presents the proposed stochastic Newton MCMC method with MAP-based Hessian, while Section 6.3.5 describes our low rank-based Hessian approximation. Section 6.4 introduces a Bayesian formulation of an ice sheet flow inverse problem, and gives expressions for adjoint-based gradient and Hessian-vector products (of the negative log posterior). In Section 6.5, we discuss the performance of the three sampling methods. Finally, in Section 6.6 we interpret the posterior distribution by visualizing marginals with respect to the eigenvectors of the covariance operator. This provides insight into the ability of the observations to infer model parameters. Section 6.7 provides concluding remarks.

6.2 Background on the infinite-dimensional Bayesian inverse problem, its consistent discretization, and characterization of the posterior

Formulating and solving the Bayesian inverse problem for an infinite-dimensional parameter field presents difficulties. First, the usual notion of a pdf is not defined since there is no Lebesgue measure in infinite dimensions. Second, the prior measure must be chosen appropriately to lead to a well-posed inverse problem and facilitate computation of the posterior. Third, the choice

of discretization must be consistent with the infinite-dimensional structure of the problem. Finally, exploring the posterior that arises upon discretization via an MCMC method is typically prohibitive due to the resulting high dimensionality of the parameter space.

In this section we formulate the Bayesian inverse problem in infinite dimensions (Section 6.2.1) in the framework of [189], which uses the Radon-Nikodym derivative and an appropriately chosen Gaussian prior that employs as covariance operator the inverse of an elliptic differential operator. In Section 6.2.2, we describe the discretization of this infinite-dimensional inverse problem in a way that is consistent with the underlying infinite-dimensional function spaces. This leads to non-standard definitions of operator adjoints. When the posterior is nearly Gaussian, its mean and covariance can be approximated by the MAP point and the inverse of the Hessian evaluated at the MAP. Inversion of the Hessian is intractable in high dimensions; Section 6.3.5 presents a low-rank approximation of the Hessian of the data misfit in order to make these Hessian computations tractable. When the posterior is not approximately Gaussian, the method of choice is often to sample it with an MCMC method and then compute sample statistics; Section 6.2.3 gives an overview of MCMC methods for sampling posteriors.

6.2.1 Bayesian formulation of infinite-dimensional inverse problems

In an inverse problem, we seek to infer the unknown (or uncertain) input parameters to a mathematical model from observations of the outputs of the

model. For ill-posed inverse problems, the uncertain parameter $m \in \mathcal{H}$ is often a heterogeneous field over a domain Ω, and $\mathcal{H}$ is typically a subset of $L^2(\Omega)$. The mathematical model is characterized by the parameter-to-observable map $\boldsymbol{f} : \mathcal{H} \to \mathbb{R}^q$, which predicts observables $\boldsymbol{y} \in \mathbb{R}^q$ corresponding to a given parameter m. Note that this map involves solution of the forward problem, typically a system of PDEs, followed by an application of an observation operator. We assume here that the observables $\boldsymbol{y}$ are finite-dimensional. Given observation data $\boldsymbol{d}^{\text{obs}} \in \mathbb{R}^q$, the solution to the inverse problem seeks parameters m such that

$$\boldsymbol{f}(m) \approx \boldsymbol{d}^{\text{obs}}$$

in a sense made precise by the Bayesian formulation described next.

The Bayesian formulation poses the inverse problem as a problem of statistical inference over parameter space. The solution of the resulting *Bayesian inverse problem* is a probability distribution that represents our belief about the correct value of the parameter. Solving the inverse problem using Bayes' approach requires specification of a *prior model*, which describes our beliefs about the parameter before any data are considered, and a *likelihood model*, which quantifies the relative probability that a candidate parameter m could have produced the observed data $\boldsymbol{d}^{\text{obs}}$.

Here we present a summary of the discussion in [33]. The prior is taken to be the Gaussian measure $\mu_0 = \mathcal{N}(m_0, \mathcal{C}_0)$ on $L^2(\Omega)$, where $m_0 \in \mathcal{H}$, and $\mathcal{C}_0$ is an appropriate covariance operator $\mathcal{C}_0$; in particular, $\mathcal{C}_0$ must be symmetric, positive, and of trace-class [189]. We choose the covariance operator to be the

inverse of an elliptic differential operator $\mathcal{A}$ that is of sufficiently high order to guarantee a well-posed Bayesian inverse problem [189]. We choose $\mathcal{A}$ to be second order differential operator[2] expressed in weak form: for $s \in L^2(\Omega)$, the solution $m = \mathcal{A}^{-1}s$ satisfies

$$\int_\Omega [a\nabla m \cdot \nabla p + bmp]\, d\boldsymbol{x} = \int_\Omega sp\, d\boldsymbol{x} \quad \text{for all } p \in H^1(\Omega), \tag{6.1}$$

with $a, b > 0$. These coefficients control the correlation length and the variance in the covariance operator $\mathcal{A}^{-1}$. Choosing for spatially dependent coefficients a and b or a tensor coefficient a allows the incorporation of further problem specific knowledge, such as spatially varying or anisotropic correlations, in the covariance operator $\mathcal{A}^{-1}$ [33].

For the likelihood model, we assume that observational uncertainty (i.e., uncertainty in $\boldsymbol{d}^{\text{obs}}$ related to measurement error) and model uncertainty (i.e., uncertainty in $\boldsymbol{f}(m)$ due to inadequacy of the forward model) are each centered, additive, and Gaussian. We combine these into a single *noise model*,

$$\boldsymbol{f}(m) = \boldsymbol{y} + \boldsymbol{\eta}, \qquad \text{with} \qquad \boldsymbol{\eta} \sim \mathcal{N}\left(\boldsymbol{0}, \boldsymbol{\Gamma}_{\text{noise}}\right),$$

where $\boldsymbol{\eta} \in \mathbb{R}^q$ is a random variable representing noise, and $\boldsymbol{\Gamma}_{\text{noise}} \in \mathbb{R}^{q\times q}$ is the noise covariance matrix. We can then express the pdf for the likelihood model

[2] The necessary order of $\mathcal{A}$ to lead to a valid covariance operator depends on the spatial dimension of the domain Ω [189]. In the example considered in Section 6.4, the inversion parameter is a one-dimensional field, and a second order differential operator is sufficient to guarantee that $\mathcal{C}_0$ is a valid covariance operator. While there is no distinction in one dimension between ordinary and partial derivatives, we choose to express $\mathcal{A}$ in the language of PDEs for notational consistency with the development in [33].

explicitly as

$$\pi_{\text{like}}(\boldsymbol{d}^{\text{obs}}|m) \propto \exp\left[-\frac{1}{2}(\boldsymbol{f}\,(m) - \boldsymbol{d}^{\text{obs}})^T \boldsymbol{\Gamma}_{\text{noise}}^{-1}(\boldsymbol{f}\,(m) - \boldsymbol{d}^{\text{obs}})\right]. \tag{6.2}$$

Bayes' theorem in infinite dimensions is expressed using the Radon-Nikodym derivative $\frac{d\mu^y}{d\mu_0}$ of the posterior measure μ^y with respect to the prior measure μ_0,

$$\frac{d\mu^y}{d\mu_0} = \frac{1}{Z}\pi_{\text{like}}(\boldsymbol{d}^{\text{obs}}|m), \tag{6.3}$$

where $Z = \int_X \pi_{\text{like}}(\boldsymbol{d}^{\text{obs}}|m)\, d\mu_0$ is a normalization constant. For technical conditions under which the posterior measure is well defined, and a discussion of the Bayes rule for probability measures on function spaces, we refer the reader to [52, 53, 189].

6.2.2 Discretization of the Bayesian inverse problem

In this section, we present a brief discussion of the finite-dimensional approximations of the prior and the posterior distributions; a lengthier discussion can be found in [33]. We start with a finite-dimensional subspace V_h of $L^2(\Omega)$ originating from a finite element discretization with continuous Lagrange basis functions $\{\phi_j\}_{j=1}^n$ [18, 188]. The approximation of the inversion parameter function $m \in L^2(\Omega)$ is then $m_h = \sum_{j=1}^n m_j\phi_j \in V_h$, where the vector of the n inversion parameters is $\boldsymbol{m} = (m_1, \ldots, m_n)^T \in \mathbb{R}^n$.

Since we postulate the prior Gaussian measure on $L^2\,(\Omega)$, the finite-dimensional space V_h inherits the L^2-inner product. Thus, inner products between nodal coefficient vectors must be weighted by a mass matrix $\boldsymbol{M} \in \mathbb{R}^{n\times n}$

to approximate the infinite-dimensional L^2-inner product. This M-weighted inner product is denoted by $\langle \cdot\,, \cdot \rangle_{\boldsymbol{M}}$, where $\langle \boldsymbol{y}, \boldsymbol{z} \rangle_{\boldsymbol{M}} = \boldsymbol{y}^T \boldsymbol{M} \boldsymbol{z}$ and $\boldsymbol{M}$ is the (symmetric positive definite) mass matrix

$$M_{ij} = \int_\Omega \phi_i(\boldsymbol{x}) \phi_j(\boldsymbol{x})\, d\boldsymbol{x}\ , \quad i, j = 1, \ldots, n.$$

To distinguish $\mathbb{R}^n$ equipped with the M-weighted inner product with the usual Euclidean space $\mathbb{R}^n$, we denote it by $\mathbb{R}^n_{\boldsymbol{M}}$.

When using the M-weighted inner product, there is a critical distinction that must be made between the matrix adjoint and the matrix transpose. For an operator $\boldsymbol{B} : \mathbb{R}^n_{\boldsymbol{M}} \to \mathbb{R}^n_{\boldsymbol{M}}$, we denote the matrix transpose by $\boldsymbol{B}^T$ with entries $(B^T)_{ij} = B_{ji}$. In contrast, the M-weighted inner product adjoint $\boldsymbol{B}^*$ satisfies, for $\boldsymbol{y}, \boldsymbol{z} \in \mathbb{R}^n$,

$$\langle \boldsymbol{B}\boldsymbol{y}, \boldsymbol{z} \rangle_{\boldsymbol{M}} = \langle \boldsymbol{y}, \boldsymbol{B}^* \boldsymbol{z} \rangle_{\boldsymbol{M}}\,,$$

which implies that $\boldsymbol{B}^*$ is given by

$$\boldsymbol{B}^* = \boldsymbol{M}^{-1} \boldsymbol{B}^T \boldsymbol{M}. \tag{6.4}$$

In the following, we also need the adjoint $\boldsymbol{V}^\diamond$ of $\boldsymbol{V} : \mathbb{R}^r \to \mathbb{R}^n_{\boldsymbol{M}}$ (for some r), where $\mathbb{R}^r$ is endowed with the Euclidean inner product. In this case, we have

$$\boldsymbol{V}^\diamond = \boldsymbol{V}^T \boldsymbol{M}, \tag{6.5}$$

since $\langle \boldsymbol{V}\boldsymbol{y}, \boldsymbol{z} \rangle_{\boldsymbol{M}} = \langle \boldsymbol{y}, \boldsymbol{V}^\diamond \boldsymbol{z} \rangle$. With these definitions, the matrix representation of the elliptic PDE operator $\mathcal{A}$ defined by (6.1) is given by $\boldsymbol{A} = \boldsymbol{M}^{-1}\boldsymbol{K} \in$

$\mathbb{R}^{n\times n}$ [33], where $\boldsymbol{K} \in \mathbb{R}^{n\times n}$ is the stiffness matrix

$$K_{ij} = \int_\Omega [a\nabla\phi_i(\boldsymbol{x}) \cdot \nabla\phi_j(\boldsymbol{x}) + b\phi_i(\boldsymbol{x})\phi_j(\boldsymbol{x})]\, d\boldsymbol{x}, \quad i,j \in \{1,\dots,n\}\,.$$

Then, the finite-dimensional approximation μ_0^h of the prior Gaussian measure μ_0 is the more familiar multivariate Gaussian with density

$$\pi_{\text{prior}}(\boldsymbol{m}) \propto \exp\left[-\frac{1}{2}\left\langle \boldsymbol{m}-\boldsymbol{m}_0, \boldsymbol{A}(\boldsymbol{m}-\boldsymbol{m}_0)\right\rangle_{\boldsymbol{M}}\right], \tag{6.6}$$

where $\boldsymbol{m}_0 \in \mathbb{R}^n$ is the discretization of the prior mean m_0. The finite-dimensional Bayes' formula, i.e.,

$$\pi_{\text{post}}(\boldsymbol{m}) := \pi_{\text{post}}(\boldsymbol{m}|\boldsymbol{d}^{\text{obs}}) \propto \pi_{\text{prior}}(\boldsymbol{m})\pi_{\text{like}}(\boldsymbol{d}^{\text{obs}}|\boldsymbol{m}), \tag{6.7}$$

where $\pi_{\text{post}}(\boldsymbol{m}|\boldsymbol{d}^{\text{obs}})$ is the density of the finite-dimensional approximation $\mu^{y,h}$ of the posterior measure μ^y, and π_{like} is the likelihood (6.2), gives the finite-dimensional posterior density explicitly as

$$\pi_{\text{post}}(\boldsymbol{m}) \propto \exp\left[-\frac{1}{2}\left\|\boldsymbol{f}(\boldsymbol{m})-\boldsymbol{d}^{\text{obs}}\right\|^2_{\boldsymbol{\Gamma}^{-1}_{\text{noise}}} - \frac{1}{2}\left\langle \boldsymbol{m}-\boldsymbol{m}_0, \boldsymbol{\Gamma}^{-1}_{\text{prior}}(\boldsymbol{m}-\boldsymbol{m}_0)\right\rangle_{\boldsymbol{M}}\right], \tag{6.8}$$

where $\boldsymbol{\Gamma}_{\text{prior}} = \boldsymbol{A}^{-1}$. Note that in (6.8) and the remainder of this paper we denote by $\boldsymbol{f}(\boldsymbol{m})$ the parameter-to-observable map evaluated at the finite element function corresponding to the parameter vector $\boldsymbol{m}$. The Bayesian solution of the inverse problem is then given by (6.8). Unfortunately, for inverse problems governed by expensive forward models and for high-dimensional parameter spaces, exploring the posterior density $\pi_{\text{post}}(\boldsymbol{m})$ is extremely challenging, since evaluation of this density at any point in parameter space requires the

solution of the forward model $\boldsymbol{f}(\boldsymbol{m})$ for the given $\boldsymbol{m}$, and a very large number of such evaluations will be required in high dimensions. Methods for exploring $\pi_{\text{post}}(\boldsymbol{m})$ that do not exploit its structure are thus impractical.

We observe that the negative log posterior density is analogous to the least squares functional that is minimized in the solution of a deterministic inverse problem. That is,

$$-\log \pi_{\text{post}}(\boldsymbol{m}) = J(\boldsymbol{m}) + \text{ const.} \tag{6.9}$$

where

$$J(\boldsymbol{m}) := \tfrac{1}{2}\|\boldsymbol{f}(\boldsymbol{m}) - \boldsymbol{d}^{\text{obs}}\|^2_{\boldsymbol{\Gamma}^{-1}_{\text{noise}}} + \frac{1}{2}\left\langle \boldsymbol{m} - \boldsymbol{m}_0, \boldsymbol{\Gamma}^{-1}_{\text{prior}}(\boldsymbol{m} - \boldsymbol{m}_0)\right\rangle_{\boldsymbol{M}}. \tag{6.10}$$

In the context of deterministic inversion, the first term in (6.10) is the data misfit term, weighted by $\boldsymbol{\Gamma}^{-1}_{\text{noise}}$, and the second term plays the role of Tikhonov regularization, which is chosen to make the inverse problem well-posed. This connection between the negative log posterior and the deterministic inverse problem cost function in (6.10) is often exploited to find an approximation of the mean of the posterior pdf by finding the point that maximizes the posterior $\pi_{\text{post}}(\boldsymbol{m})$, or equivalently minimizes the cost function $J(\boldsymbol{m})$. This so-called maximum a posterior (MAP) point is equal to the mean when the parameter-to-observable map $\boldsymbol{f}(\boldsymbol{m})$ is linear in the parameters $\boldsymbol{m}$ and the noise and prior models are Gaussian. When the Gaussian-linear conditions are not satisfied, obviously the MAP point only approximates the mean, the quality of this approximation depending on the degree of nonlinearity. Moreover, under

these Gaussian-linear conditions, the posterior $\pi_{\text{post}}(\boldsymbol{m})$ is Gaussian with mean given by the MAP point, and covariance given by the inverse of the Hessian matrix of the cost function $J(\boldsymbol{m})$ [33, 192].

6.2.3 Exploring the posterior

As implied above, when the parameter-to-observable map is nonlinear, the posterior $\pi_{\text{post}}(\boldsymbol{m})$ generally is non-Gaussian, and cannot be represented by its mean and covariance. Thus it must be characterized by other means. This can be extremely challenging for PDE-based inverse problems, since evaluating the posterior (6.8) at any point in parameter space involves solving the forward PDEs, and many such evaluations are anticipated for the high-dimensional parameter spaces that stem from discretization of infinite-dimensional inverse problems.

The method of choice for exploring the posterior pdf is the Metropolis-Hastings (M-H) MCMC method [104, 149, 174, 193], which employs a given proposal probability density $q(\boldsymbol{m}_k, \boldsymbol{y})$ at each sample point $\boldsymbol{m}_k$ in parameter space to generate a proposed sample point $\boldsymbol{y} \in \mathbb{R}^n$. Once generated, the M-H criterion chooses to either accept or reject the proposed sample point, and repeats from the new point, thereby generating a chain of samples $\{\boldsymbol{m}_k\}_{k=1,\cdots}$ from the posterior density $\pi_{\text{post}}(\boldsymbol{m})$. Algorithm 7 presents pseudo-code for the M-H MCMC method.

Critical to the success of M-H MCMC is the choice of the proposal density $q(\boldsymbol{m}_k, \boldsymbol{y})$. Observe that if $q(\boldsymbol{m}_k, \boldsymbol{y}) = \pi_{\text{post}}(\boldsymbol{y})$, the M-H algorithm would

Algorithm 7 Metropolis-Hastings MCMC algorithm to sample the pdf π

Choose initial parameters $\boldsymbol{m}_0$
Compute $\pi(\boldsymbol{m}_0)$
for $k = 0, \ldots, N-1$ **do**
  Draw sample $\boldsymbol{y}$ from the proposal density $q(\boldsymbol{m}_k, \cdot)$
  Compute $\pi(\boldsymbol{y})$
  Compute $\alpha_k(\boldsymbol{y}) = \min\left\{1, \frac{\pi(\boldsymbol{y})q(\boldsymbol{y},\boldsymbol{m}_k)}{\pi(\boldsymbol{m}_k)q(\boldsymbol{m}_k,\boldsymbol{y})}\right\}$
  Draw $u \sim \mathcal{U}([0,1])$
  if $u < \alpha_k(\boldsymbol{y})$ **then**
    Accept: Set $\boldsymbol{m}_{k+1} = \boldsymbol{y}$
  else
    Reject: Set $\boldsymbol{m}_{k+1} = \boldsymbol{m}_k$
  end if
end for

accept every sample with probability 1; however, this defeats the purpose, because we would not know how to sample from this choice of proposal: the whole point of appealing to MCMC is that we cannot draw a sample directly from $\pi_{\text{post}}(\boldsymbol{y})$.

Instead, a common choice for the proposal is the isotropic Gaussian,

$$q^{\text{RWMH}}(\boldsymbol{m}_k, \boldsymbol{y}) = \frac{1}{(2\pi)^{n/2}} \exp[-\tfrac{1}{2}(\|\boldsymbol{m}_k - \boldsymbol{y}\|)^2].$$

The resulting method is known as Random Walk Metropolis-Hastings (RWMH). This proposal density is easy to sample, but it can lead to poor MCMC performance due to the mismatch between the proposal and posterior densities. The challenge is to come up with a proposal that at least locally reflects the behavior of the target posterior density and at the same time is easy to sample. Satisfying these two requirements becomes increasingly difficult with increasing parameter dimension. This will be the subject of the next section.

6.3 A modified stochastic Newton MCMC method

In [143], we introduced a so-called stochastic Newton MCMC method that featured a Gaussian proposal constructed from the local gradient vector and local Hessian matrix (of the negative log posterior). To make the construction of the proposal tractable, we employed adjoint-based methods to compute the gradient and Hessian, which amount to a pair of forward/adjoint PDE solves for the gradient and for each column of the Hessian. Moreover, to make the Hessian computation scalable with respect to parameter dimension, we use matrix-free methods to construct low-rank approximations of the data misfit component of the Hessian, which often has a rapidly-decaying spectrum reflecting the ill-posedness of the inverse problem [73, 143]. With these features, the stochastic Newton method is able to handle inverse problems with hundreds to thousands of parameters; its efficiency increases with decreasing nonlinearity of the parameter-to-observable map and with decreasing information content of the data. We denote this original form of the stochastic Newton MCMC method (i.e., with dynamically changing Hessian) as *SN*.

Unfortunately, SN becomes prohibitive for very large-scale problems, because it requires recomputation of the Hessian at each sample point. Despite the use of efficient adjoint-based matrix-free Hessian-vector products to find the low-rank approximations of the data misfit component of the Hessian, we still need $O(2r)$ linearized forward/adjoint PDE solves to compute it, where r is the effective rank. When r is large—as is the case for high-dimensional problems, for which the observations are highly informative about the param-

eters and, hence, the data misfit Hessian has a high effective rank—we must find alternatives to computing the Hessian at each sample point.

Here we propose a modified stochastic Newton MCMC method that employs a (low-rank approximation-based) Hessian that is computed once and for all at the MAP point, and reused for each proposal. This modification, to which we refer as *stochastic Newton MCMC with MAP-based Hessian (SNMAP)*, employs a locally-computed gradient in the Gaussian proposal, but evaluates the Hessian in that Gaussian at the MAP point. Before describing SNMAP, we begin with a brief summary of the proposal construction for the original stochastic Newton MCMC method.

6.3.1 Stochastic Newton MCMC with dynamically changing Hessian (SN)

The stochastic Newton MCMC method employs a local Gaussian approximation of the target posterior pdf. This is done by constructing, about a given point $\boldsymbol{m}_k$, a local quadratic approximation $\tilde{J}_k(\boldsymbol{m})$ of the negative log posterior $J(\boldsymbol{m})$ (given in (6.10)), i.e.,

$$\tilde{J}_k(\boldsymbol{m}) := J(\boldsymbol{m}_k) + \langle \boldsymbol{g}_k, \boldsymbol{m} - \boldsymbol{m}_k \rangle_{\boldsymbol{M}} + \frac{1}{2} \langle \boldsymbol{m} - \boldsymbol{m}_k, \boldsymbol{H}_k(\boldsymbol{m} - \boldsymbol{m}_k) \rangle_{\boldsymbol{M}} . \quad (6.11)$$

Here, $\boldsymbol{g}$ and $\boldsymbol{H}$ are the gradient vector and Hessian matrix of $J(\boldsymbol{m})$, respectively, and $\boldsymbol{g}_k := \boldsymbol{g}(\boldsymbol{m}_k) \in \mathbb{R}^n$ and $\boldsymbol{H}_k := \boldsymbol{H}(\boldsymbol{m}_k) \in \mathbb{R}^{n \times n}$. Rearranging terms,

$$\tilde{J}_k(\boldsymbol{m}) = \frac{1}{2} \left\langle \boldsymbol{m} - \boldsymbol{m}_k + \boldsymbol{H}_k^{-1} \boldsymbol{g}_k, \boldsymbol{H}_k(\boldsymbol{m} - \boldsymbol{m}_k + \boldsymbol{H}_k^{-1} \boldsymbol{g}_k) \right\rangle_{\boldsymbol{M}} + \text{const.}$$

To obtain the proposal density q^{SN} for stochastic Newton MCMC (with dynamically changing Hessian), we take the exponential of the negative of $\tilde{J}_k(\boldsymbol{m})$, and compute the scaling factor to make it a proper pdf. This leads to

$$q^{\text{SN}}(\boldsymbol{m}_k, \boldsymbol{y}) = \frac{\det \boldsymbol{H}_k^{1/2}}{(2\pi)^{n/2}} \exp\left(-\frac{1}{2}\left\langle \boldsymbol{y} - \boldsymbol{m}_k + \boldsymbol{H}_k^{-1}\boldsymbol{g}_k, \boldsymbol{H}_k(\boldsymbol{y} - \boldsymbol{m}_k + \boldsymbol{H}_k^{-1}\boldsymbol{g}_k)\right\rangle_{\boldsymbol{M}}\right), \tag{6.12}$$

which is a Gaussian with mean $\boldsymbol{m}_k - \boldsymbol{H}_k^{-1}\boldsymbol{g}_k$ and covariance matrix $\boldsymbol{H}_k^{-1}$. Note that at a local minimum, $\boldsymbol{H}_k$ is positive semi-definite and at an arbitrary point $\boldsymbol{y}$, $\boldsymbol{H}_k$ can be indefinite. To ensure that (6.12) defines a proper pdf, we discard negative eigenvalues of the data misfit component of $\boldsymbol{H}_k$ and, hence, replace $\boldsymbol{H}_k$ with a modified positive definite Hessian. We also note that the backward proposal $q^{\text{SN}}(\boldsymbol{y}, \boldsymbol{m}_k)$, needed for the M-H acceptance probability α_k, is computed using the Hessian and gradient evaluated at $\boldsymbol{y}$. In summary, the SN step at each MCMC iteration draws a proposed sample $\boldsymbol{y}$ from the proposal $q^{\text{SN}}(\boldsymbol{m}_k, \boldsymbol{y})$, which is then subject to the accept/reject framework of the M-H MCMC Algorithm 7. The SN proposal is illustrated in Figure 6.1 (top left).

6.3.2 Stochastic Newton MCMC with MAP-based Hessian (SN-MAP)

As stated above, the original form of the stochastic Newton MCMC method becomes prohibitive for very large-scale problems, because it requires recomputation of the Hessian of $J(\boldsymbol{m})$ (whose inverse is needed to construct the Gaussian proposal) at each sample point. Therefore, we avoid recomputing this Hessian by the following modification: we first find the MAP point

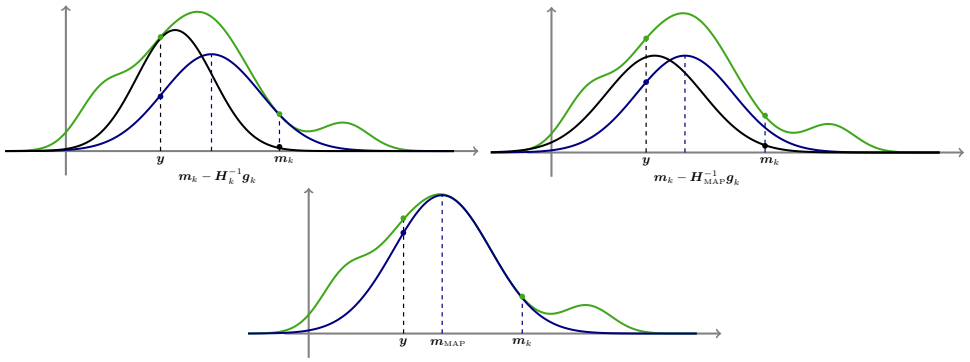

Figure 6.1: Illustration of proposals for the three Hessian-based methods: stochastic Newton MCMC with dynamically-computed Hessian (top left); stochastic Newton MCMC with MAP-based Hessian (top right); and independence sampler with MAP-based Hessian (bottom). The green curve depicts the true posterior density, $\pi_{\text{post}}(\boldsymbol{m})$; the blue curve displays the forward proposal density, $q(\boldsymbol{m}_k, \boldsymbol{y})$; and the black curve shows the backward proposal density, $q(\boldsymbol{y}, \boldsymbol{m}_k)$. The green, blue, and black dotted lines indicate the points at which the posterior, the backward, and the forward proposals, respectively, are evaluated.

and compute the Hessian there, and then use this MAP-based Hessian for all proposals. The gradient is still computed at each sample point. Hence, the proposal $q^{\text{SNMAP}}(\boldsymbol{m}_k, \boldsymbol{y})$ for SNMAP is given by replacing the Hessian in (6.12) with the Hessian evaluated at the MAP point, $\boldsymbol{H}_{\text{MAP}}$. This leads to

$$q^{\text{SNMAP}}(\boldsymbol{m}_k, \boldsymbol{y}) \propto \exp\left(-\frac{1}{2}\left\langle \boldsymbol{y} - \boldsymbol{m}_k + \boldsymbol{H}_{\text{MAP}}^{-1}\boldsymbol{g}_k, \boldsymbol{H}_{\text{MAP}}(\boldsymbol{y} - \boldsymbol{m}_k + \boldsymbol{H}_{\text{MAP}}^{-1}\boldsymbol{g}_k)\right\rangle_{\boldsymbol{M}}\right), \tag{6.13}$$

which is a Gaussian with mean $\boldsymbol{m}_k - \boldsymbol{H}_{\text{MAP}}^{-1}\boldsymbol{g}_k$ and covariance matrix $\boldsymbol{H}_{\text{MAP}}^{-1}$. Note that the scaling factor is not necessary in (6.13), since for proposals with MAP-based Hessians, the scaling factors are constant and thus they cancel when computing the acceptance probability α_k in Algorithm 7. The SNMAP proposal is illustrated in Figure 6.1 (top right). Note that the SNMAP proposal (6.13) can also be understood as a preconditioned Langevin MCMC proposal [190] with a MAP-based Hessian preconditioner.

Avoiding SN's Hessian recomputation at each sample point results in substantial computational savings, since, as will be made explicit in Section 6.3.5, computing the Hessian typically requires a number of forward/adjoint PDE solves on the order of the effective rank of (a properly preconditioned) Hessian of the data misfit term in the negative log posterior. Once SNMAP has computed the Hessian at the MAP point, the only cost per sample is a pair of forward/adjoint PDE solves to compute the gradient. However, this may result in a deterioration in the acceptance rate, since the Gaussian proposal employs local gradient information but a global Hessian, and thus may not fully capture local curvature information of the posterior if the curvature is changing rapidly

(as may happen in a highly nonlinear parameter-to-observable map). However, the fact that the proposal q^{SNMAP} changes less from sample to sample compared to q^{SN} can also have a positive effect on the acceptance probability and the chain convergence. In Section 6.5 we conduct numerical experiments on a specific Bayesian inverse problem to assess whether this tradeoff is profitable.

6.3.3 Independence sampling with a MAP point-based Gaussian proposal (ISMAP)

As seen in the above SNMAP modification of the stochastic Newton MCMC method, freezing the Hessian at the MAP point avoids Hessian recomputation and results in substantial savings. However, the gradient is still recomputed at each sample point, motivated by the desire to construct a Gaussian proposal that captures some local information, as well as the fact that the gradient is far cheaper to compute than (a low rank-approximation-based) Hessian.

One can go one step further and shed the need to compute local gradient information by defining an independence sampler that takes the proposal to be a Gaussian centered at the MAP point, using the Hessian computed at the MAP as the inverse covariance, and neglecting the gradient (since it vanishes at the MAP). This method has been suggested previously in the subsurface flow inversion literature [156,157]; here, we refer to it as *ISMAP*. Since ISMAP, like SN and SNMAP, makes use of Hessian information, a fair assessment of SNMAP vis-á-vis SN should include comparisons to ISMAP as well. Therefore,

we next provide a description of the ISMAP proposal as well.

The proposal density q^{ISMAP} is obtained by taking $\boldsymbol{m}_k$ in (6.12) as the MAP point $\boldsymbol{m}_{\text{MAP}}$ (which means that $\boldsymbol{g}_k$ is zero) and, as with SNMAP, replacing the Hessian at $\boldsymbol{m}_k$ with the Hessian evaluated at the MAP point, $\boldsymbol{H}_{\text{MAP}}$. This leads to

$$q^{\text{ISMAP}}(\boldsymbol{m}_{\text{MAP}}, \boldsymbol{y}) \propto \exp\left(-\frac{1}{2}\left\langle \boldsymbol{y} - \boldsymbol{m}_{\text{MAP}}, \boldsymbol{H}_{\text{MAP}}(\boldsymbol{y} - \boldsymbol{m}_{\text{MAP}})\right\rangle_{\boldsymbol{M}}\right), \tag{6.14}$$

which is a Gaussian with mean $\boldsymbol{m}_{\text{MAP}}$ and covariance matrix $\boldsymbol{H}_{\text{MAP}}^{-1}$. We note that the proposal q^{ISMAP} is independent of the current sample point, and thus does not change during the sampling process. The ISMAP proposal is illustrated in Figure 6.1 (bottom).

We note that ISMAP not only avoids Hessian recomputation at each sample point (as with SNMAP) but also avoids computing the gradient; thus, its cost—once the MAP-based Hessian is determined—is a forward PDE solve at each sample point. However, this additional approximation over SNMAP has the potential to lead to additional deterioration of the acceptance rate. Note that one advantage of ISMAP is that, since the proposal is constant, the samples can all be precomputed offline or in parallel, after which they can be subjected (sequentially) to the M-H accept/reject criterion in Algorithm 7.

Finally, we remark that if the posterior itself is a Gaussian, the three Hessian-based methods described above collapse to the same method. As such, they all sample from the true posterior with probability 1 at every step,

resulting in an acceptance rate of 100% and posterior samples that are independent [143].

6.3.4 Relation to Newton's method for optimization

Recall that the stochastic Newton MCMC method (in particular SN) uses, as a proposal, the local quadratic approximation $\tilde{J}_k(\boldsymbol{m})$ of the negative log posterior $J(\boldsymbol{m})$ about the current sample point $\boldsymbol{m}_k$. The minimizer of $\tilde{J}_k(\boldsymbol{m})$ is given by $\boldsymbol{m}_k - \boldsymbol{H}_k^{-1}\boldsymbol{g}_k$, where $\boldsymbol{H}_k$ and $\boldsymbol{g}_k$ are the Hessian and the gradient of $J(\boldsymbol{m})$ evaluated at $\boldsymbol{m}_k$, respectively. Note that $-\boldsymbol{H}_k^{-1}\boldsymbol{g}_k$ is the classical Newton optimization step. A proposal point drawn from the local Gaussian approximation of the posterior with mean $\boldsymbol{m}_k - \boldsymbol{H}_k^{-1}\boldsymbol{g}_k$ and covariance $\boldsymbol{H}_k^{-1}$ is thus

$$\boldsymbol{y} = \boldsymbol{m}_k - \boldsymbol{H}_k^{-1}\boldsymbol{g}_k + \boldsymbol{H}_k^{-1/2}\tilde{\boldsymbol{n}}, \tag{6.15}$$

where $\tilde{\boldsymbol{n}} = \boldsymbol{M}^{-1/2}\boldsymbol{n}$ is a random sample from a Gaussian with zero mean and identity covariance matrix in $\mathbb{R}^n_{\boldsymbol{M}}$, and $\boldsymbol{n} \in \mathbb{R}^n$ is a random sample from the standard normal density in $\mathbb{R}^n$. Iterating the stochastic Newton MCMC method without the random term amounts to the classical Newton method from nonlinear optimization, which converges to the MAP point (or another stationary point of $J(\cdot)$).

Since SNMAP reuses the Hessian at the MAP point (i.e., it is held constant throughout the sampling process), proposal points are computed as

in (6.15), but with $\boldsymbol{H}_k$ replaced by the Hessian at the MAP point $\boldsymbol{H}_{\text{MAP}}$, i.e.,

$$\boldsymbol{y} = \boldsymbol{m}_k - \boldsymbol{H}_{\text{MAP}}^{-1}\boldsymbol{g}_k + \boldsymbol{H}_{\text{MAP}}^{-1/2}\tilde{\boldsymbol{n}}, \tag{6.16}$$

with $\tilde{\boldsymbol{n}}$ as above. We note that if the random term is neglected, SNMAP reduces to an $\boldsymbol{H}_{\text{MAP}}$-preconditioned steepest descent method.

For completeness, let us show how proposals from the independence sampler with MAP-based Gaussian, ISMAP, are computed. With $\tilde{\boldsymbol{n}}$ defined as above, the proposed point is found as

$$\boldsymbol{y} = \boldsymbol{m}_{\text{MAP}} + \boldsymbol{H}_{\text{MAP}}^{-1/2}\tilde{\boldsymbol{n}}. \tag{6.17}$$

Note that the right hand side in (6.17) is independent of $\boldsymbol{m}_k$, and hence the designation "independence sampler."

6.3.5 Efficient operations with the Hessian via low-rank approximation

Up to this point, we have described the three Hessian-based MCMC methods (SN, SNMAP, and ISMAP) in terms of the Hessian matrix of the negative log posterior. Indeed, examination of the form of the three proposal densities (6.12), (6.13), and (6.14), as well as the expressions for the samples from the proposals (6.15), (6.16), and (6.17), reveals that the following operations with the Hessian are required: action of the Hessian on a vector; action of the inverse Hessian on a vector; action of the inverse of the square root of the Hessian on a vector; and determinant of the square root of the Hessian (the determinant is required only for SN).

Unfortunately, explicitly computing the Hessian requires as many (linearized) forward PDE solves as there are parameters; for large-scale problems, these computations are prohibitive. Thus, we need efficient algorithms for the operations with the Hessian summarized above. In this section, we briefly describe previous work that employs low-rank approximations of the data misfit portion of the Hessian, preconditioned by the prior covariance, to execute all of the above operations with the Hessian at a cost (measured in forward PDE solves) that is independent of the parameter dimension [33, 73, 143]. The discussion below is in terms of a generic Hessian, $\boldsymbol{H}$; this can refer to the Hessian at any point in parameter space, including the MAP point.

The Hessian of the negative log posterior $J(\boldsymbol{m})$ in (6.10) can be written as the sum of the Hessian of the data misfit term, $\boldsymbol{H}_{\text{misfit}}$, and the inverse of the prior covariance $\boldsymbol{\Gamma}_{\text{prior}}^{-1}$. If we consider a decomposition of the prior such that $\boldsymbol{\Gamma}_{\text{prior}} = \boldsymbol{L}\boldsymbol{L}^*$, then

$$\boldsymbol{H} = \boldsymbol{H}_{\text{misfit}} + \boldsymbol{\Gamma}_{\text{prior}}^{-1} = \boldsymbol{H}_{\text{misfit}} + \boldsymbol{L}^{-*}\boldsymbol{L}^{-1} = \boldsymbol{L}^{-*}(\boldsymbol{L}^*\boldsymbol{H}_{\text{misfit}}\boldsymbol{L} + \boldsymbol{I})\boldsymbol{L}^{-1}. \quad (6.18)$$

Here, the data misfit Hessian $\boldsymbol{H}_{\text{misfit}}$ is given by

$$\boldsymbol{H}_{\text{misfit}} := \boldsymbol{F}^\natural \boldsymbol{\Gamma}_{\text{noise}}^{-1} \boldsymbol{F} \; + \; \text{second order terms}$$

where $\boldsymbol{F}$ is the Jacobian matrix of the parameter-to-observable map $\boldsymbol{f}(\boldsymbol{m})$, $\boldsymbol{F}^\natural := \boldsymbol{M}^{-1}\boldsymbol{F}^T$ is its (properly weighted) adjoint, and the second order terms involve second derivatives of $\boldsymbol{f}(\boldsymbol{m})$ with respect to $\boldsymbol{m}$. Notwithstanding the form of the second order terms, the expression above suggests that the Hessian

of the data misfit involves the solution of linearized forward and adjoint PDE problems. This will be seen explicitly for the target ice sheet inverse problem described in Section 6.4.

We begin by describing the computation of the application of the inverse Hessian to a vector in order to compute the Newton step $\boldsymbol{H}^{-1}\boldsymbol{g}$. From (6.18), we obtain

$$\boldsymbol{H}^{-1}\boldsymbol{g} = \boldsymbol{L}(\boldsymbol{L}^*\boldsymbol{H}_{\text{misfit}}\boldsymbol{L} + \boldsymbol{I})^{-1}\boldsymbol{L}^*\boldsymbol{g}, \tag{6.19}$$

and thus we require the inverse of $(\boldsymbol{L}^*\boldsymbol{H}_{\text{misfit}}\boldsymbol{L} + \boldsymbol{I})$. Since for ill-posed inverse problems, observations typically inform only a limited number of eigenvectors of the parameter field, the spectrum of the data misfit Hessian often decays rapidly (see for example, [29,30,31] for the inverse scattering case). In addition, the prior is often smoothing, in which case left and right preconditioning of the data misfit Hessian by the square root of the prior, $\boldsymbol{L}$, enhances the decay of the eigenvalues. Thus, the prior-preconditioned data misfit Hessian, $\boldsymbol{L}^*\boldsymbol{H}_{\text{misfit}}\boldsymbol{L}$, can typically be well approximated by a low rank matrix, and this can be exploited to enable efficient computations with the Hessian. To construct the low-rank approximation of the prior-preconditioned data misfit Hessian, we seek a matrix-free method (since $\boldsymbol{H}$ cannot be formed explicitly) that requires just Hessian-vector products; crucially, the number of Hessian-vector products must be of the order of the effective rank, r, of the prior-preconditioned data misfit Hessian, as opposed to the parameter dimension, n. Note that each Hessian-vector product can be formed efficiently at the cost of a single pair of linearized forward/adjoint PDE solves (this will be seen explicitly for the ice

sheet flow problem in Section 6.4.5).

The Lanczos eigenvalue algorithm meets the requirements outlined above, and we use it to construct an r-dimensional low-rank approximation for the prior-preconditioned data misfit Hessian, i.e., $\boldsymbol{L}^*\boldsymbol{H}_{\text{misfit}}\boldsymbol{L} \approx \boldsymbol{V}_r\boldsymbol{\Lambda}_r\boldsymbol{V}_r^\diamond$, where $\boldsymbol{V}_r \in \mathbb{R}^{n\times r}$ contains r eigenvectors of the prior-preconditioned data misfit Hessian corresponding to the r largest eigenvalues $\lambda_i, i = 1,\dots,r$, $\boldsymbol{\Lambda}_r = \text{diag}(\lambda_1,\dots,\lambda_r) \in \mathbb{R}^{r\times r}$, and $\boldsymbol{V}_r^\diamond \in \mathbb{R}^{r\times n}$ denotes the adjoint defined in (6.5). The rank r approximation can typically be formed in a number of Hessian-vector products that is slightly larger than r, which amounts to approximately r forward/adjoint pairs of linearized PDE solves all containing the same PDE operator or its adjoint (this presents an opportunity to employ an effective PDE preconditioner, since it will be amortized over r PDE forward/adjoint solves.) Once the low-rank approximation has been constructed, the product of this approximate Hessian with a vector can then be formed by successively applying $\boldsymbol{V}_r$ and $\boldsymbol{V}_r^\diamond$ to vectors, each application amounting to r inner products. The cost of this linear algebra is negligible relative to the PDE solves needed to form the low-rank approximation.

Moreover, using the Sherman-Morrison-Woodbury formula [89] in combination with expressing the prior-preconditioned data misfit Hessian as the sum of a low rank term and a reminder, we can write the inverse Hessian as

$$(\boldsymbol{L}^*\boldsymbol{H}_{\text{misfit}}\boldsymbol{L} + \boldsymbol{I})^{-1} = \boldsymbol{I} - \boldsymbol{V}_r\boldsymbol{D}_r\boldsymbol{V}_r^\diamond + \mathcal{O}\left(\sum_{i=r+1}^{n} \frac{\lambda_i}{\lambda_i + 1}\right), \tag{6.20}$$

where $\boldsymbol{D}_r := \text{diag}(\lambda_1/(\lambda_1+1),\dots,\lambda_r/(\lambda_r+1)) \in \mathbb{R}^{r\times r}$. As can be seen from

the form of the remainder term above, to obtain an accurate low rank approximation of $\boldsymbol{H}^{-1}$, we can neglect eigenvectors corresponding to eigenvalues that are small compared to 1. Therefore,

$$\boldsymbol{H}^{-1}\boldsymbol{g} \approx \boldsymbol{L}(\boldsymbol{I}-\boldsymbol{V}_r\boldsymbol{D}_r\boldsymbol{V}_r^{\diamond})\boldsymbol{L}^*\boldsymbol{g} = \boldsymbol{L}\Big\{\boldsymbol{V}_r\Big[(\boldsymbol{\Lambda}_r+\boldsymbol{I}_r)^{-1}-\boldsymbol{I}_r\Big]\boldsymbol{V}_r^{\diamond}+\boldsymbol{I}\Big\}\boldsymbol{L}^*\boldsymbol{g}. \quad (6.21)$$

The expression on the right side of (6.21) can be used to efficiently apply the square-root inverse Hessian to a vector $\boldsymbol{x}$, as needed for drawing samples from a Gaussian distribution with covariance $\boldsymbol{H}^{-1}$. Namely,

$$\boldsymbol{H}^{-1/2}\boldsymbol{x} \approx \boldsymbol{L}\Big\{\boldsymbol{V}_r\Big[(\boldsymbol{\Lambda}_r+\boldsymbol{I}_r)^{-1/2}-\boldsymbol{I}_r\Big]\boldsymbol{V}_r^{\diamond}+\boldsymbol{I}\Big\}\boldsymbol{x}. \quad (6.22)$$

By a direct computation using the adjoint definitions (6.4) and (6.5), it can be verified that $\boldsymbol{H}^{-1}\boldsymbol{x} = \boldsymbol{H}^{-1/2}(\boldsymbol{H}^{-1/2})^*\boldsymbol{x}$. Finally, the determinant of the square-root Hessian can be computed efficiently from

$$\det(\boldsymbol{H}^{1/2}) = (\det \boldsymbol{L})^{-1}\prod_{i=1}^{r}(\lambda_i+1)^{1/2}. \quad (6.23)$$

In summary, once the low-rank approximation of the data misfit Hessian has been constructed, all of the operations with the Hessian described above (and required by the three Hessian-based methods) can be carried out using only inner products and vector sums, without recourse to PDE solves. These linear algebra operations are negligible relative to the PDE solves needed for the low-rank approximation, and thus the dominant cost of these methods is $O(r)$ forward/adjoint PDE solves needed for the low-rank approximation. As mentioned above, for ill-posed inverse problems (including the ice sheet flow

inverse problem studied below), the prior-preconditioned data misfit Hessian is a compact operator with rapidly-decaying eigenvalues, so that $r \ll n$. Moreover, when the dominant eigenvectors of the prior-preconditioned data misfit Hessian are spatially smooth, r is independent of the parameter dimension n and the observation dimension q.

6.3.6 Comparison of computational cost of ISMAP, SNMAP, and SN

The stochastic Newton MCMC methods and the independence sampler with MAP-based Gaussian all use the low-rank approximation and fast operations with the Hessian described in the previous section. However, they differ markedly in how frequently they recompute the low-rank approximation of the prior-preconditioned data misfit Hessian, which as mentioned above is by far the dominant cost relative to the linear algebra.

Let us now characterize the cost per MCMC sample for each of the three Hessian-based methods described in Section 6.3, measured in number of (forward or adjoint) PDE solves. The independence sampling method (ISMAP) requires just a single evaluation of the parameter-to-observable map per sample, which amounts to a single (nonlinear) forward PDE solve per sample. The stochastic Newton MCMC method with dynamically changing Hessian (SN) requires for each sample a nonlinear forward PDE solve, a (linear) adjoint PDE solve for the gradient computation, and approximately $2r$ linearized PDE solves to construct the rank r approximation of the prior-preconditioned

data misfit Hessian. Finally, the cost per sample for stochastic Newton MCMC with MAP-based Hessian (SNMAP) is one nonlinear forward PDE solve and one adjoint PDE solve, since SNMAP recomputes the gradient at each sample point. Depending on whether the forward problem is linear or nonlinear and stationary or time dependent, and depending on whether the linearized PDEs are solved by direct factorization (which permits reuse of the factors within the low-rank approximation) or iteratively (which permits reuse of only the preconditioner), the number of PDE solves per sample translates differently into computational time per sample. Thus, the metric we use to compare the performance of these three Hessian-based methods to each other in Section 6.5 is the number of linearized PDE solves required by each method.

6.4 Application to the inversion of basal boundary conditions in ice flow problems

In the remainder of this paper, we apply the methods discussed in Section 6.3 to an inverse problem in ice dynamics, in which we seek to find a statistical description of the uncertain basal sliding coefficient field from pointwise velocity observations at the surface of the moving mass of ice. In this section, we summarize the physics describing the dynamics of ice flows, present the two-dimensional problem used to exercise our methods, and the prior distribution and the likelihood for the Bayesian inverse problem. We also give expressions of the gradient and the Hessian-vector product of the negative log posterior function using adjoint ice flow equations and describe

the discretization of these equations.

6.4.1 The dynamics of ice flow

We model the flow of ice as a non-Newtonian, viscous, incompressible, isothermal fluid [92, 115, 142, 162]. The balance of mass and linear momentum in a domain $\Omega \subset \mathbb{R}^d$ of dimension $d = 2$ or $d = 3$ state that

$$\boldsymbol{\nabla} \cdot \boldsymbol{u} = 0 \quad \text{in } \Omega, \tag{6.24a}$$

$$-\boldsymbol{\nabla} \cdot \boldsymbol{\sigma}_{\boldsymbol{u}} = \rho \boldsymbol{g} \quad \text{in } \Omega, \tag{6.24b}$$

where $\boldsymbol{u}$ denotes the velocity vector, $\boldsymbol{\sigma}_{\boldsymbol{u}}$ the stress tensor, ρ the density of the ice, and $\boldsymbol{g}$ gravity. The stress, $\boldsymbol{\sigma}_{\boldsymbol{u}}$, can be decomposed as $\boldsymbol{\sigma}_{\boldsymbol{u}} = \boldsymbol{\tau}_{\boldsymbol{u}} - \boldsymbol{I}p$, where $\boldsymbol{\tau}_{\boldsymbol{u}}$ is the deviatoric stress tensor, p the pressure, and $\boldsymbol{I}$ the unit tensor. We employ a constitutive law for ice that relates stress and strain rate tensors by Glen's flow law [87],

$$\boldsymbol{\tau}_{\boldsymbol{u}} = 2\eta(\boldsymbol{u})\dot{\boldsymbol{\varepsilon}}_{\boldsymbol{u}}, \text{ with } \eta(\boldsymbol{u}) = \frac{1}{2} A^{-\frac{1}{n}} \, \dot{\boldsymbol{\varepsilon}}_{\text{II}}^{\frac{1-n}{2n}}, \tag{6.24c}$$

where η is the effective viscosity, $\dot{\boldsymbol{\varepsilon}}_{\boldsymbol{u}} = \frac{1}{2}(\boldsymbol{\nabla}\boldsymbol{u} + \boldsymbol{\nabla}\boldsymbol{u}^T)$ the strain rate tensor, $\dot{\boldsymbol{\varepsilon}}_{\text{II}} = \frac{1}{2}\text{tr}(\dot{\boldsymbol{\varepsilon}}_{\boldsymbol{u}}^2)$ its second invariant, $n \geq 1$ Glen's flow law exponent, and A the temperature-dependent flow rate factor (here taken as constant in isothermal ice).

At the base Γ_{b} of the ice sheet, one commonly assumes non-penetrating normal boundary conditions and a linear sliding law for the tangential components i.e. [162]

$$\boldsymbol{u} \cdot \boldsymbol{n} = 0, \;\; \boldsymbol{T}\boldsymbol{\sigma}_{\boldsymbol{u}}\boldsymbol{n} + \exp(\beta)\boldsymbol{T}\boldsymbol{u} = \boldsymbol{0}, \tag{6.24d}$$

where $\beta = \beta(\boldsymbol{x})$ is the log basal sliding coefficient field, and $\boldsymbol{T} := \boldsymbol{I} - \boldsymbol{n} \otimes \boldsymbol{n}$ the projection onto the tangential plane. Here, "$\otimes$" represents the tensor (or outer) product defined by $(\boldsymbol{a} \otimes \boldsymbol{b})\boldsymbol{c} = \boldsymbol{a}\boldsymbol{b} \cdot \boldsymbol{c}$, $\boldsymbol{n}$ is the outward normal vector, and $\boldsymbol{I}$ is the second order unit tensor. Together with appropriate boundary conditions on $\partial\Omega \setminus \Gamma_b$, (6.24) represents an accepted model for the flow of ice sheets and glaciers. Note that the Robin coefficient field $\exp(\beta)$, which relates tangential velocity to tangential traction, subsumes several complex physical phenomena such as the frictional behavior of the ice sheet, the roughness of the bedrock and hydrological phenomena. It does not itself represent a physical parameter and is highly uncertain. Our target is to infer the log sliding coefficient field β, which in the following we simply refer to as sliding coefficient field, within a Bayesian inversion approach. In the next section, we specify the ice flow model problem used to study the efficiency of our algorithms and to interpret results of Bayesian inversions.

6.4.2 The Arolla test problem

We use a two-dimensional test problem taken from the Ice Sheet Model Intercomparison Project for Higher-Order Ice Sheet Models (ISMIP-HOM) benchmark study [163]. The domain Ω, which is based on data from the Haut Glacier d'Arolla is shown in Figure 6.2. Together with the basal boundary condition (6.24d), on the top boundary Γ_t we assume the traction-free condition

$$\boldsymbol{\sigma}_{\boldsymbol{u}} \boldsymbol{n} = \boldsymbol{0} \text{ on } \Gamma_t. \tag{6.24e}$$

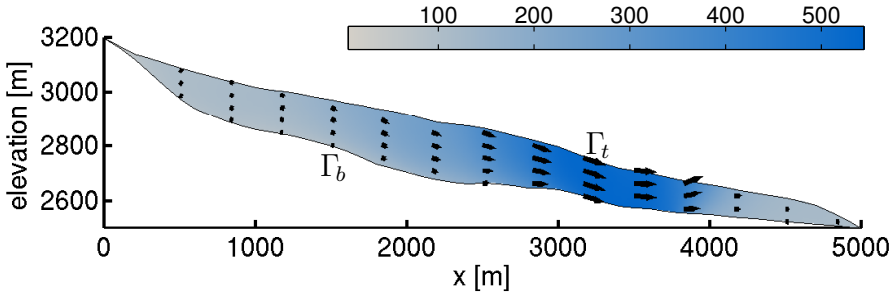

Figure 6.2: The longitudinal profile of Haut Glacier d'Arolla from the ISMIP-HOM benchmark collection [163]. This profile follows a flowline of 5 km length with a grid spacing of 100 m. The arrows represent the flow field obtained by solving (6.24) with the basal sliding coefficient field given by (6.25).

The driving force in the Stokes equations (6.24) is the gravity $\rho \boldsymbol{g} = (0, -\rho g \cos\theta)$, where $\rho = 910$ kg/m^3 is the ice density, and $g = 9.81$ m/s^2 is the gravitational constant. The Glen's flow-law exponent parameter is $n = 3$, and the rate factor is assumed constant as $A = 10^{-16}$ Pa^{-n}a^{-1}, where "Pa" and "a" are units of Pascals and years, respectively [163].

As reference basal sliding coefficient field, which is also used to generate synthetic observations as described in the next section, we choose

$$\beta_{\text{true}}(x) = \ln \begin{cases} 1000 + 1000 \sin\left(\frac{2\pi x}{5000}\right) & \text{if } 0 \le x < 3750, \\ 1000 \left(16 - \frac{x}{250}\right) & \text{if } 3750 \le x < 4000, \\ 1000 & \text{if } 4000 \le x < 5000. \end{cases} \tag{6.25}$$

The flow field corresponding to the basal sliding coefficient field (6.25) are shown in Figure 6.2.

6.4.3 The likelihood

The likelihood function expresses the probability that a candidate set of parameters reproduces the observations $\boldsymbol{d}^{\text{obs}}$. To specify the likelihood function, we denote by $\boldsymbol{u}(\beta)$ the solution of the Stokes equation with basal sliding coefficient field $\beta(\boldsymbol{x})$, and by $\mathcal{B}$ the observation operator, which restricts the flow solution to ten measurement points on the right of the top surface Γ_t, i.e., lower part of the glacier, with x-coordinates uniformly distributed in $[2500, 5000]$. Thus, the parameter-to-observable map is $\boldsymbol{f}(\beta) = \mathcal{B}\boldsymbol{u}(\beta)$.

The observations $\boldsymbol{d}^{\text{obs}}$ are synthetically generated by solving the ice flow Stokes equations (6.24) with basal sliding coefficient field β_{true} as specified in (6.25), restricting the resulting flow solution $\boldsymbol{u}(\beta_{\text{true}})$ using the observation operator and adding additive Gaussian noise $\boldsymbol{\epsilon}$, i.e., $\boldsymbol{d}^{\text{obs}} = \mathcal{B}\boldsymbol{u}(\beta_{\text{true}}) + \boldsymbol{\epsilon}$. Each component of the noise vector $\boldsymbol{\epsilon}$ is i.i.d. with standard deviation $\bar{\sigma}_{\text{noise}}$ for the horizontal flow components and with $\bar{\bar{\sigma}}_{\text{noise}}$ for the vertical flow components. Adding noise mitigates the "inverse crime," which occurs when synthetic observations are used in an inversion and the same numerical method is employed in both, the synthetization of the observations and in the inverse problem solution [117]. The likelihood function is then given by

$$\pi_{\text{like}}(\boldsymbol{d}^{\text{obs}}|\boldsymbol{\beta}) \propto \exp\left[-\frac{1}{2}\left\|\mathcal{B}\boldsymbol{u}(\beta) - \boldsymbol{d}^{\text{obs}}\right\|^2_{\boldsymbol{\Gamma}^{-1}_{\text{noise}}}\right], \tag{6.26}$$

where the noise covariance matrix $\boldsymbol{\Gamma}_{\text{noise}}$ is diagonal with the entries $\bar{\sigma}_{\text{noise}}$ and $\bar{\bar{\sigma}}_{\text{noise}}$ for the horizontal and vertical components, respectively. To understand the effect of the noise level on the performance of the three Hessian-based

sampling methods and on the uncertainty in the reconstruction, we consider two problems based on the noise level in the observations:

- **Problem 1:** $\bar{\sigma}_{\text{noise}} = 62$ for the horizontal flow components and with $\bar{\bar{\sigma}}_{\text{noise}} = 10$ for the vertical flow components;
- **Problem 2:** $\bar{\sigma}_{\text{noise}} = 18$ for the horizontal flow components and with $\bar{\bar{\sigma}}_{\text{noise}} = 3$ for the vertical flow components.

6.4.4 The choice of prior

We specify the Gaussian prior by giving its mean β_0 and its covariance via the elliptic operator $\mathcal{A}$ discussed in Section 6.2. Since the bottom surface of the Arolla geometry is a "curved" surface, the prior is defined in terms of the surface Laplacian (also called the Laplace-Beltrami operator). Using the projection $\boldsymbol{T}$ onto the tangential plane as defined above, $\boldsymbol{\nabla}_{\Gamma_b} = \boldsymbol{T}\boldsymbol{\nabla}$ is the tangential gradient, $\boldsymbol{\nabla}_{\Gamma_b}\cdot$ is the tangential divergence, and $\boldsymbol{\nabla}_{\Gamma_b} \cdot \boldsymbol{\nabla}_{\Gamma_b}$ is the Laplace-Beltrami operator [22, 59, 64]. Thus, we define $\mathcal{A}$ as the differential operator

$$-\boldsymbol{\nabla}_{\Gamma_b} \cdot (a\boldsymbol{\nabla}_{\Gamma_b}\beta) + b\beta = s \quad \text{in } \Gamma_b, \tag{6.27a}$$

$$(a\boldsymbol{\nabla}_{\Gamma_b}\beta) \cdot \boldsymbol{\nu} = 0 \quad \text{on } \partial\Gamma_b, \tag{6.27b}$$

where $\boldsymbol{\nu}$ denotes the outward unit normal on $\partial\Gamma_b$. The finite-dimensional representation of the prior inverse is $\boldsymbol{\Gamma}_{\text{prior}}^{-1} = \boldsymbol{M}^{-1}\boldsymbol{K}$, where $\boldsymbol{M}$ and $\boldsymbol{K}$ are the corresponding surface mass and surface stiffness matrices, respectively. In

our model problems, we use the parameters $a = 10^{-2}$ and $b = 10^2$. With these parameters, the standard deviation of the Green's function corresponding to the prior (i.e., the correlation length) is roughly 5% of the total length of the glacier.

6.4.5 Gradient and Hessian of the negative log posterior

The Hessian-based sampling methods presented in Section 6.3 rely on the availability of gradients and Hessian-vector products of the negative log posterior. The derivation of these derivatives is complicated by the fact that the parameter-to-observable map involves the solution of the ice flow equations. In this section, we give expressions for the efficient computation of gradients and Hessian-vector products using adjoint equations. For a more detailed presentation of derivative computation using adjoints we refer to the PDE-constrained optimization monographs [23, 112, 195] and to [167] for the ice flow dynamics setting.

The gradient of the negative log posterior can be found by requiring that variations of a Lagrangian function with respect to the forward velocity and pressure $(\boldsymbol{u}, p)$ and an adjoint velocity and pressure $(\boldsymbol{v}, q)$ vanish. Variations with respect to β then result in the following strong form of the gradient $\mathcal{G}$:

$$\mathcal{G}(\beta) := \exp(\beta)\boldsymbol{T}\boldsymbol{u} \cdot \boldsymbol{T}\boldsymbol{v} + \mathcal{A}(\beta - \beta_0). \tag{6.28}$$

Here, the velocity $\boldsymbol{u}$ is obtained by solving the *forward Stokes problem* (6.24) for given β, and the adjoint velocity $\boldsymbol{v}$ is obtained by solving the following

adjoint Stokes problem for given β and for $\boldsymbol{u}$ satisfying (6.24):

$$\boldsymbol{\nabla} \cdot \boldsymbol{v} = 0 \qquad \text{in } \Omega, \tag{6.29a}$$

$$-\boldsymbol{\nabla} \cdot \boldsymbol{\sigma_v} = \boldsymbol{0} \qquad \text{in } \Omega, \tag{6.29b}$$

$$\boldsymbol{\sigma_v} \boldsymbol{n} = -\mathcal{B}^* \Gamma_{\text{noise}}^{-1} (\mathcal{B}\boldsymbol{u} - \boldsymbol{d}^{\text{obs}}) \qquad \text{on } \Gamma_t, \tag{6.29c}$$

$$\boldsymbol{v} \cdot \boldsymbol{n} = 0, \ \boldsymbol{T}\boldsymbol{\sigma_v}\boldsymbol{n} + \exp(\beta)\boldsymbol{T}\boldsymbol{v} = \boldsymbol{0} \qquad \text{on } \Gamma_b, \tag{6.29d}$$

where the adjoint stress $\boldsymbol{\sigma_v}$ is given by

$$\boldsymbol{\sigma_v} := 2\eta(\boldsymbol{u}) \left(\mathsf{I} + \frac{1-n}{n} \frac{\dot{\boldsymbol{\varepsilon}}_{\boldsymbol{u}} \otimes \dot{\boldsymbol{\varepsilon}}_{\boldsymbol{u}}}{\dot{\boldsymbol{\varepsilon}}_{\boldsymbol{u}} : \dot{\boldsymbol{\varepsilon}}_{\boldsymbol{u}}} \right) \dot{\boldsymbol{\varepsilon}}_{\boldsymbol{v}} - \boldsymbol{I} q,$$

and I is the fourth order identity tensor.

The action of the Hessian operator evaluated at a sliding coefficient field β onto a direction $\hat{\beta}$ is given by

$$\mathcal{H}(\beta)(\hat{\beta}) := \mathcal{A}\hat{\beta} + \exp(\beta)(\hat{\beta}\boldsymbol{T}\boldsymbol{u} \cdot \boldsymbol{T}\boldsymbol{v} + \boldsymbol{T}\hat{\boldsymbol{u}} \cdot \boldsymbol{T}\boldsymbol{v} + \boldsymbol{T}\boldsymbol{u} \cdot \boldsymbol{T}\hat{\boldsymbol{v}}), \tag{6.30}$$

where the *incremental forward velocity/pressure* $(\hat{\boldsymbol{u}}, \hat{p})$ satisfy the *incremental forward Stokes problem*,

$$\boldsymbol{\nabla} \cdot \hat{\boldsymbol{u}} = 0 \qquad \text{in } \Omega, \tag{6.31a}$$

$$-\boldsymbol{\nabla} \cdot \boldsymbol{\sigma}_{\hat{\boldsymbol{u}}} = \boldsymbol{0} \qquad \text{in } \Omega, \tag{6.31b}$$

$$\boldsymbol{\sigma}_{\hat{\boldsymbol{u}}} \boldsymbol{n} = \boldsymbol{0} \qquad \text{on } \Gamma_t, \tag{6.31c}$$

$$\hat{\boldsymbol{u}} \cdot \boldsymbol{n} = 0, \ \boldsymbol{T}\boldsymbol{\sigma}_{\hat{\boldsymbol{u}}}\boldsymbol{n} + \exp(\beta)\boldsymbol{T}\hat{\boldsymbol{u}} = -\hat{\beta}\exp(\beta)\boldsymbol{T}\boldsymbol{u} \qquad \text{on } \Gamma_b, \tag{6.31d}$$

with $\boldsymbol{\sigma}_{\hat{\boldsymbol{u}}} := 2\eta(\boldsymbol{u})\left(\mathsf{I} + \frac{1-n}{n}\frac{\dot{\boldsymbol{\varepsilon}}_{\boldsymbol{u}} \otimes \dot{\boldsymbol{\varepsilon}}_{\boldsymbol{u}}}{\dot{\boldsymbol{\varepsilon}}_{\boldsymbol{u}} : \dot{\boldsymbol{\varepsilon}}_{\boldsymbol{u}}}\right)\dot{\boldsymbol{\varepsilon}}_{\hat{\boldsymbol{u}}} - \boldsymbol{I}\hat{p}$, and the *incremental adjoint velocity/pressure* $(\hat{\boldsymbol{v}}, \hat{q})$ satisfy the *incremental adjoint Stokes problem*,

$$\boldsymbol{\nabla} \cdot \hat{\boldsymbol{v}} = 0 \qquad \text{in } \Omega, \tag{6.32a}$$

$$-\boldsymbol{\nabla} \cdot \boldsymbol{\sigma}_{\hat{\boldsymbol{v}}} = -\boldsymbol{\nabla} \cdot \tau_{\hat{\boldsymbol{u}}} \qquad \text{in } \Omega, \tag{6.32b}$$

$$\boldsymbol{\sigma}_{\hat{\boldsymbol{v}}}\boldsymbol{n} = -\mathcal{B}^* \Gamma_{\text{noise}}^{-1} \mathcal{B}\hat{\boldsymbol{u}} - \tau_{\hat{\boldsymbol{u}}} \quad \text{on } \Gamma_t, \tag{6.32c}$$

$$\hat{\boldsymbol{v}} \cdot \boldsymbol{n} = 0,\ \boldsymbol{T}\boldsymbol{\sigma}_{\hat{\boldsymbol{v}}}\boldsymbol{n} + \exp(\beta)\boldsymbol{T}\hat{\boldsymbol{v}} = -\boldsymbol{T}\tau_{\hat{\boldsymbol{u}}}\boldsymbol{n} \qquad \text{on } \Gamma_b, \tag{6.32d}$$

with $\boldsymbol{\sigma}_{\hat{\boldsymbol{v}}} := 2\eta(\boldsymbol{u})\left(\mathsf{I} + \frac{1-n}{n}\frac{\dot{\boldsymbol{\varepsilon}}_{\boldsymbol{u}} \otimes \dot{\boldsymbol{\varepsilon}}_{\boldsymbol{u}}}{\dot{\boldsymbol{\varepsilon}}_{\boldsymbol{u}} : \dot{\boldsymbol{\varepsilon}}_{\boldsymbol{u}}}\right)\dot{\boldsymbol{\varepsilon}}_{\hat{\boldsymbol{v}}} - \boldsymbol{I}\hat{q}$, and $\tau_{\hat{\boldsymbol{u}}} = 2\eta(\boldsymbol{u})\Psi\dot{\boldsymbol{\varepsilon}}_{\hat{\boldsymbol{u}}}$, where

$$\Psi = (1 + \frac{1-n}{n}\dot{\boldsymbol{\varepsilon}}_{\boldsymbol{u}} : \dot{\boldsymbol{\varepsilon}}_{\boldsymbol{u}})\mathsf{I} + \frac{1-n}{n}\left[\frac{\dot{\boldsymbol{\varepsilon}}_{\boldsymbol{u}} \otimes \dot{\boldsymbol{\varepsilon}}_{\boldsymbol{u}}}{\dot{\boldsymbol{\varepsilon}}_{\boldsymbol{u}} : \dot{\boldsymbol{\varepsilon}}_{\boldsymbol{u}}} + 2\frac{\dot{\boldsymbol{\varepsilon}}_{\boldsymbol{u}} \otimes \dot{\boldsymbol{\varepsilon}}_{\boldsymbol{v}}}{\dot{\boldsymbol{\varepsilon}}_{\boldsymbol{u}} : \dot{\boldsymbol{\varepsilon}}_{\boldsymbol{u}}} + \frac{1-3n}{n}\frac{\dot{\boldsymbol{\varepsilon}}_{\boldsymbol{u}} \otimes \dot{\boldsymbol{\varepsilon}}_{\boldsymbol{u}}}{(\dot{\boldsymbol{\varepsilon}}_{\boldsymbol{u}} : \dot{\boldsymbol{\varepsilon}}_{\boldsymbol{u}})^2}\right].$$

In these expressions, $\dot{\boldsymbol{\varepsilon}}_{\hat{\boldsymbol{u}}}$ and $\dot{\boldsymbol{\varepsilon}}_{\hat{\boldsymbol{v}}}$ are defined analogously to $\dot{\boldsymbol{\varepsilon}}_{\boldsymbol{u}}$ and $\dot{\boldsymbol{\varepsilon}}_{\boldsymbol{v}}$.

To summarize, the computational cost (measured in the number of linearized Stokes solves, which represent the dominant cost) of the gradient evaluation is n_{ls} forward linearized Stokes solves for the nonlinear forward problem (6.24) (where n_{ls} is the number of Newton iterations required by the nonlinear solver to converge), and one linear adjoint solve for (6.29). Each computation of the Hessian-vector product (6.30) requires two linearized Stokes solves, namely the solution of (6.31) and (6.32).

6.4.6 Discretization and solvers

We discretize the domain Ω with 260 triangular mesh elements, and use Taylor-Hood finite elements (i.e., linear elements for pressure and quadratic

elements for the velocity components which leads to 4714 degrees of freedom for the velocity field and 659 for the pressure) for the forward and adjoint Stokes problems as well as their incremental counterparts. The uncertain sliding coefficient field β is discretized using linear elements with 139 unknowns, i.e., parameters for the inverse problem. We ensure that the state and parameter fields are sufficiently resolved by comparing the solutions computed on different meshes. All Stokes systems are solved using a direct factorization method. The cost of the Stokes matrix factorization is amortized across the adjoint solve and the incremental forward and adjoint solves in all CG iterations needed in each Newton iteration; when the factorization is available, only triangular solutions are required at each CG iteration, gradient computation or the application of the Hessian to a vector.

6.5 Performance of algorithms

The primary goal of this section is to compare the performance of the sampling methods presented in Section 6.3 for the Bayesian inverse problem described in Section 6.4. We start with a discussion on the computation of the MAP point in Section 6.5.1, and study the approximation of prior-preconditioned data misfit Hessians—and thus covariance matrices—using low rank ideas (Section 6.5.2). In Section 6.5.3, we present a systematic comparison of the three Hessian-based sampling methods (ISMAP, SN, and SNMAP) presented in Section 6.3.

6.5.1 Computation of the MAP point

For the computation of the MAP point, we apply an adjoint-based inexact Newton method to solve the nonlinear least-squares optimization problem (6.9). Starting with an initial guess for the basal sliding coefficient field β (we use the prior mean $\beta_0 \equiv \ln(1000)$), Newton's method iteratively updates this parameter based on successive quadratic approximations of the negative log posterior functional $J(\cdot)$, using the expressions for the first and second derivatives presented in Section 6.4.5. Since the conjugate gradient method is used to solve the Newton linearization, the method does not require assembled Hessian matrices but only Hessian-vector products. For a more complete presentation of this optimization method to compute MAP points for ice sheet model problems, we refer to [167].

We discuss the performance of the optimization algorithm for the computation of the MAP point for Problem 2 as defined in Section 6.4.3. On the right in Figure 6.3, we show the "truth" sliding coefficient field, which is used to generate the synthetic surface velocity observations. Also shown is the MAP point, i.e., the solution of (6.9). In the upper part of the glacier the MAP point follows the prior mean since observations are only available in the lower half of the glacier (i.e., the right part of the domain).

To compute the MAP point, 8 (outer) Newton iterations were necessary to decrease the nonlinear residual by a factor of 10^5. In each of these outer Newton iterations, the nonlinear Stokes equation has to be solved, for which we use an (inner) Newton method. These inner Newton solves are also terminated

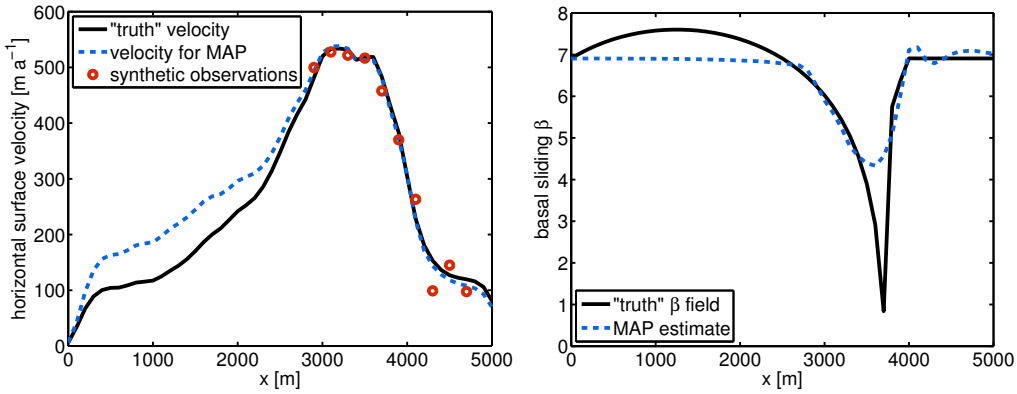

Figure 6.3: Left: The horizontal surface velocity obtained by solving the forward problem using the "truth" sliding coefficient field (solid line) and the synthetic pointwise observations (circles), generated by adding 1.5% Gaussian random noise to this surface velocity. The horizontal velocity corresponding to the MAP point is shown by the dashed line. Right: "Truth" sliding coefficient field (solid line) and MAP point (dashed line).

after the residual is decreased by a factor of 10^5, which takes an average of 12 iterations, each amounting to a linearized Stokes solve. In addition to the nonlinear Stokes solve, each (outer) Newton iteration requires computation of the gradient and of several Hessian-vector products. Summing over all 8 (outer) Newton iterations, 32 conjugate gradient iterations—and thus 32 Hessian-vector products—are required. In total, the computation of the MAP point amounts to 208 linear(ized) Stokes solves.

6.5.2 Low-rank approximation of the prior-preconditioned data misfit Hessian

The computational feasibility of Hessian-based sampling for large-scale Bayesian inverse problems critically relies on low-rank approximations for the data misfit Hessian. Thus, we study the numerical rank of the prior-preconditioned data misfit Hessian for various points in the parameter space. Figure 6.4 shows a logarithmic plot of the spectra of the prior-preconditioned data misfit Hessians at the 21 MCMC chain starting points discussed in Section 6.5.3. Note that all spectra decay rapidly. As seen in (6.20), an accurate low-rank approximation of the inverse Hessian can be obtained by neglecting eigenvalues that are small compared to 1. Thus, retaining 15–20 eigenvectors appears to be sufficient for any point from the posterior distribution.

In our sampling runs, we thus use $r = 20$ eigenvectors for the low-rank approximation of the prior-preconditioned data misfit Hessian. We note that the cost of obtaining this low-rank approximation, measured in the number of Stokes solves, is $2(r + l)$, where $r + l$ is the number of Lanczos iterations. Here, $l \geq 0$ iterations are used to ensure the accurate computation of the most significant eigenvalues/eigenvectors (we use $l = 5$). We discard any negative eigenvalues to guarantee that the low-rank approximation is positive semi-definite.

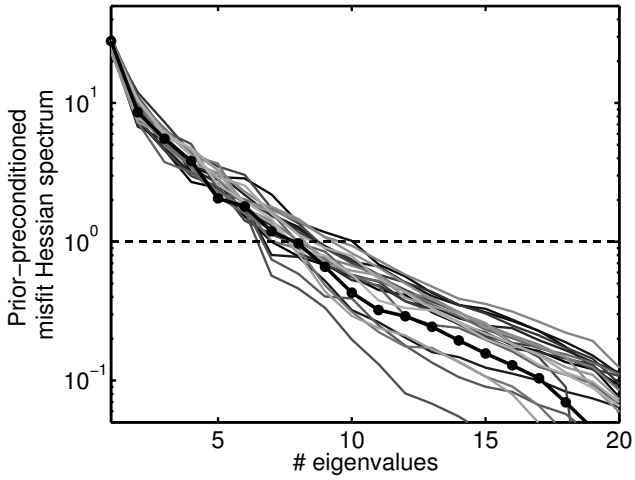

Figure 6.4: Logarithmic plot of the spectra of prior-preconditioned data misfit Hessians computed at the MAP point (black line with dots) and at 21 points distributed over the support of the posterior (gray lines). The horizontal line for $\lambda = 1$ shows the reference value for the truncation of the spectrum of the prior-preconditioned data misfit Hessian.

6.5.3 Performance of proposed stochastic Newton MCMC method with MAP-based Hessian

In this section, we compare the performance of the proposed SNMAP method (stochastic Newton MCMC with MAP-based Hessian) with SN (the original stochastic Newton MCMC method with dynamically-computed Hessian) and with ISMAP (independence sampler with MAP-based Gaussian) for both ice flow inverse problems introduced in Section 6.4.3. For each method, 21 MCMC chains are computed using a common set of 21 initial points. These points are selected from an MCMC chain with 25,000 samples initialized at the MAP point. From this chain, these 21 initial points are chosen to approximately maximize the minimum pairwise distances between points, so that the

resulting set is distributed quasi-uniformly over the support of the posterior distribution. This ensures that the initial points are over-dispersed with respect to the posterior, which is important for the convergence diagnostics used to compare the different MCMC methods.

In Table 6.1, we summarize convergence diagnostics and MCMC chain statistics averaged over 21 chains (excluding the MPSRF which is a multi-chain diagnostic). To compare the different MCMC methods, in the second column we provide the multivariate potential scale reduction factor (MPSRF) diagnostic [24]. This diagnostic compares averaged properties of individual sample chains with properties of the pooled sample chain. When these properties are similar, we infer that each of the individual sample chains has converged. The closer the MPSRF is to 1, the better converged the individual sample chains are.

It is well known in Monte Carlo methods that the variance in the estimate decays as $1/N$ when averaging over N i.i.d. samples. However, MCMC samples are not independent, and in general we observe that averaging over N samples from an MCMC chain reduces the variance in the estimate by a factor of only τ/N, where $\tau > 1$ is the integrated autocorrelation time (IAT) [174], given by

$$\tau = 1 + 2\sum_{s=1}^{\infty} \rho(s). \tag{6.33}$$

Here, $\rho(s)$ is the usual autocorrelation function for a lag $s > 0$. In practice, $\rho(s)$ is noisy when computed from a finite length sample chain, and thus we estimate τ by the maximum value obtained by truncating the summation in

(6.33). The autocorrelation is defined for a scalar quantity, and we report in column three the IAT corresponding to the sliding coefficient field at the point $x = 3450$. In the fourth column, we report the effective sample size (ESS) defined as N/τ, the number of independent samples that would be required for the same variance reduction as obtained from the MCMC chain.

The fifth column shows the mean squared jump distance (MSJ), which provides an indication of how well the MCMC chain is mixing. This metric is defined for a single MCMC chain with samples $\boldsymbol{m}_0, \ldots, \boldsymbol{m}_N$ as

$$\text{MSJ} := \frac{1}{N} \sum_{k=0}^{N-1} \|\boldsymbol{m}_{k+1} - \boldsymbol{m}_k\|_{\boldsymbol{M}}^2. \tag{6.34}$$

In general, a larger mean square jump distance indicates faster mixing of the MCMC chain, and tends to result in better chain convergence.

Finally, we address the question of greatest interest with regard to computational efficiency: "Given an MCMC algorithm, how much computational work is required to obtain an independent sample?". Column seven reports the total number of linearized Stokes solves required to obtain a single independent sample, and column eight reports the total wall-clock time for these solves.

We summarize the following observations from Table 6.1:

(a) The number of independent samples is about one order of magnitude larger for Problem 1 than for Problem 2, suggesting that the posterior distribution for Problem 2 is more difficult to sample.

Table 6.1: Multivariate potential scale reduction factor (**MPSRF**), integrated autocorrelation time (**IAT**), effective sample sample size (**ESS**), mean squared jump distance (**MSJ**), acceptance rate (**AR**), number of (linearized) Stokes solves per independent sample (**SPIS**), and the average wallclock time per independent sample (**TPIS**). We compare the performance obtained with the independence sampler with MAP-based Gaussian (ISMAP), the stochastic Newton MCMC method with MAP-based Hessian (SNMAP), and the stochastic Newton MCMC method with dynamic Hessian (SN) for two problems with different noise levels (e.g., $\bar{\sigma}_{\text{noise}} = 62$ and $\bar{\bar{\sigma}}_{\text{noise}} = 10$ for Problem 1, and $\bar{\sigma}_{\text{noise}} = 18$ and $\bar{\bar{\sigma}}_{\text{noise}} = 3$ for Problem 2). We use 21 MCMC chains, each with 25,000 samples, hence the total number of samples is 525,000. The dimension of the discretized basal sliding coefficient field, i.e., the number of parameters, is 139.

	MPSRF	**IAT**	**ESS**	**MSJ**	**AR (%)**	**SPIS**	**TPIS (s)**
Problem 1							
ISMAP	1.210	253	2075	1456	41	2783	139
SNMAP	1.001	6	84004	1390	40	72	4
SN	1.073	125	4032	565	17	1375	69
Problem 2							
ISMAP	1.507	435	1207	280	9	4350	218
SNMAP	1.045	80	6563	190	6	960	48
SN	1.348	600	875	64	2	8400	420

(b) SNMAP leads to the best MPSRF values for both problems, suggesting the fastest convergence with respect to the number of samples. As a consequence, the largest effective sample size is achieved using SNMAP. Note that this holds even though ISMAP yields larger acceptance rates and mean squared jump distances.

(c) SNMAP also requires the smallest number of forward solves per independent sample. For Problem 1, SNMAP is more efficient than SN by a factor

of about 20, and than ISMAP by a factor of almost 40. For Problem 2, SNMAP is more efficient by factors of about 10 and 5 than SN and ISMAP, respectively.

(d) Surprisingly, SN performs worse than SNMAP, even with respect to the number of samples. This is despite the fact that it uses a better local approximation of the posterior. We attribute this to the mismatch in the local Hessians at different points, which increases the asymmetry between the forward and backward proposals $q(\boldsymbol{m}_k, \boldsymbol{y})$ and $q(\boldsymbol{y}, \boldsymbol{m}_k)$, thus increasing the variability of the acceptance probability $\alpha_k(\boldsymbol{y})$ (see Algorithm 7).

We have also applied Delayed Rejection Adaptive Metropolis (DRAM) sampling [95] to explore the posterior distribution. We found it to be far from convergence after 1,000,000 samples. We attribute this to the high dimensional parameter space and the lack of information about the problem structure in the sampling process. In the next section, we focus on visualization and interpretation of the posterior distribution.

6.6 Analysis and interpretation of the solution of the Bayesian inverse problem

Visualization and interpretation for high-dimensional posterior distributions is a difficult task. In this section, we highlight techniques motivated by the structure of the Bayesian inverse problem to guide our analysis. First, in Section 6.6.1, we present visualizations of the posterior in the physical coor-

dinate basis, which provides intuition about the solution at particular points or regions of the domain. Then, in Section 6.6.2, we shift our perspective to eigenvectors of the posterior covariance (approximated using the Hessian at the MAP point), which can be classified into groups according to their contributions from the observation data and the prior. The qualitative features of each group provide insight into the posterior distribution. Finally, in Section 6.6.3, we visualize one- and two-dimensional marginal distributions of the full posterior distribution with respect to these eigenvectors.

The results discussed in this section are for Problem 2 as defined in Section 6.4.3, i.e., the problem with smaller data noise. The approximation of the posterior pdf is based on samples generated by the SNMAP method, and kernel density estimation is used to visualize the one- and two-dimensional marginal probability density functions.

6.6.1 Point marginals and samples from the posterior

In Figure 6.5, we present (one-dimensional) marginals of the prior and posterior distributions with respect to physical points in the domain. We refer to these marginals as point marginals. The probability density for each point marginal is visualized in gray scale along a vertical column at each point, with higher probability density indicated by darker shading. Because each point marginal is computed independently, the point marginal density values at neighboring points are not necessarily related, and thus any spatial correlation structure present in the distribution is neglected by this visualization. For

this reason, we overlay a few samples from each distribution to provide some indication of the spatial correlation structure.

This visualization provides some useful observations for our problem. In the unobserved part of the domain (the upper part of the glacier), the point marginals of the posterior are similar to those of the prior; our beliefs about the basal sliding coefficient field in this region are unchanged from the prior. On the contrary, in the region where observation data are available, we find the variance to be decreased significantly (i.e., we are more certain about the sliding coefficient field in this region), and in some regions most of the probability mass is shifted in the posterior compared to the prior; the evidence from the observation data has overwhelmed our prior beliefs in this region. Finally, while spatial correlation structure is difficult to infer from the limited number of overlayed samples, note that the average width of the variations appears unchanged from the prior to the posterior in both, the parts of the glacier with and without observations. We interpret this as insufficient observational evidence to update our beliefs about the width of spatial variations.

6.6.2 Classification of posterior covariance eigenvectors

In this section, we classify the eigenvectors of the posterior covariance into groups according to their contributions from the observation data and prior, and subsequently use this classification to gain insight into the posterior distribution. While it is common to order eigenvectors by ascending or

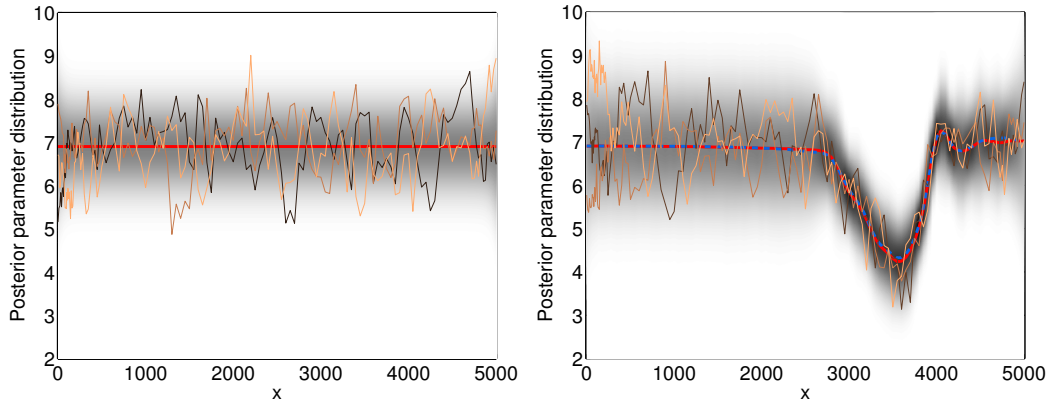

Figure 6.5: Shown in gray scale are the one-dimensional point marginals of the prior (left) and the posterior (right) probability distributions, with higher probability density indicated by darker shading. Point marginals are computed and plotted independently along a vertical line at each point, where the gray shaded area corresponds to a 95% confidence interval. To give an indication of spatial correlation, samples from the prior and the posterior are shown (in different shades of brown). Also shown are the prior and posterior mean (in red), and the MAP point of the posterior (in blue). We recall that the dimension of the discretized basal sliding coefficient field, i.e., the number of parameters, is 139.

descending eigenvalues, this choice is poorly adapted to our purposes since it unpredictably interleaves data-influenced eigenvectors with prior-influenced eigenvectors. We therefore propose a general technique for sorting eigenvectors that groups them naturally.

To characterize the influence of the observations and prior on the eigenvectors, consider the Rayleigh quotients of the data misfit Hessian and of the inverse of the prior, i.e.,

$$r_m^i = \frac{\langle \boldsymbol{v}_i, \boldsymbol{H}_{\text{misfit}} \boldsymbol{v}_i \rangle_{\boldsymbol{M}}}{\langle \boldsymbol{v}_i, \boldsymbol{v}_i \rangle_{\boldsymbol{M}}}, \quad r_p^i = \frac{\left\langle \boldsymbol{v}_i, \boldsymbol{\Gamma}_{\text{prior}}^{-1} \boldsymbol{v}_i \right\rangle_{\boldsymbol{M}}}{\langle \boldsymbol{v}_i, \boldsymbol{v}_i \rangle_{\boldsymbol{M}}}, \tag{6.35}$$

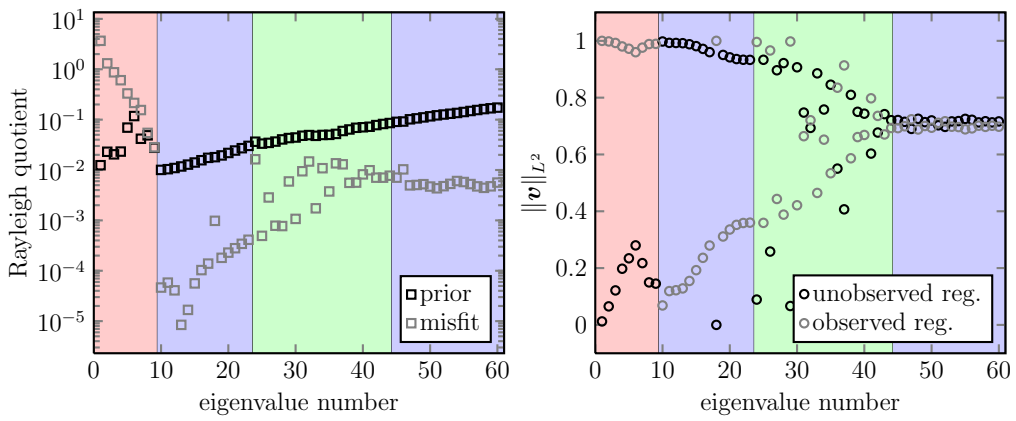

Figure 6.6: Left: Semilogarithmic plot of the Rayleigh quotients of the data misfit Hessian and of the inverse prior covariance as defined in (6.35). Right: Norm of the eigenvectors in the lower and upper parts of the glacier, i.e., in the region with and without observations, respectively.

for $i = 1, \ldots, n$, where $\boldsymbol{v}_i$ is the i-th eigenvector of the inverse posterior covariance (approximated by the Hessian at the MAP point). Because the eigenvalue λ^i associated with $\boldsymbol{v}_i$ is simply the sum of r_m^i and r_p^i, these Rayleigh quotients quantify the contributions from the observation data and prior. We then order the eigenvectors according to the difference of the squared Rayleigh coefficients $d^i := (r_m^i)^2 - (r_p^i)^2$. Large positive values of d^i correspond to eigenvectors that are most informed by the data, whereas large negative values correspond to directions most informed by the prior. We note that there are several reasonable choices for d^i; we find that our choice best groups eigenvectors with similar qualitative features. The sorted Rayleigh quotients for the data misfit Hessian and for the inverse of the prior are presented in the left plot in Figure 6.6. A selection of these eigenvectors is shown in Figure 6.7.

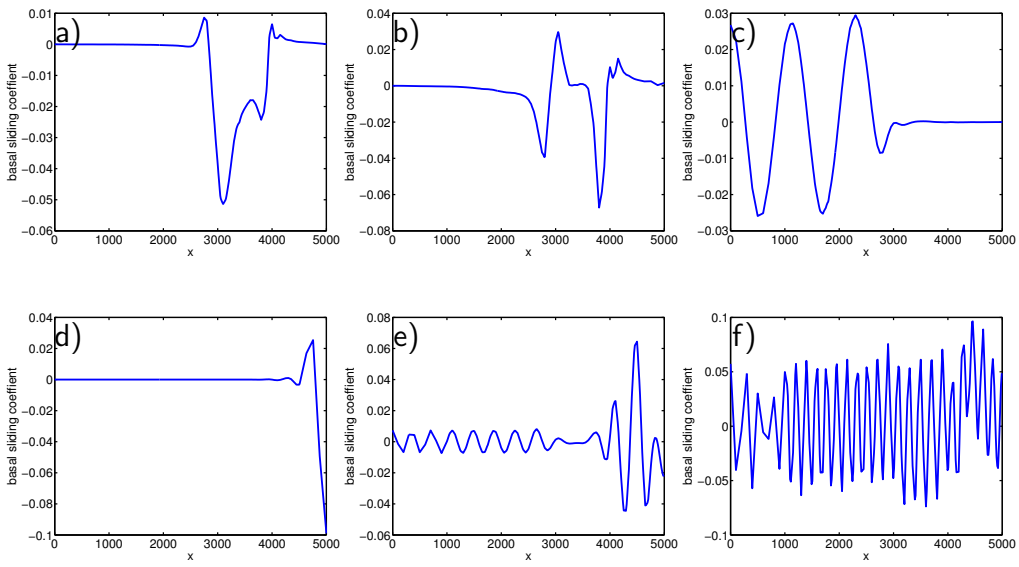

Figure 6.7: Eigenvectors of the Hessian corresponding to the 1st, 3rd, 14th, 18th, 26th and 55th eigenvalues are shown in figures a–f, respectively. Note that different eigenvectors are concentrated in different parts of the domain and that eigenvectors corresponding to smaller eigenvalues are more oscillatory.

Next, we study the qualitative features of these eigenvectors. Since the lower half of the glacier contains observation points and the upper half does not, we can also characterize these eigenvectors by determining whether the eigenvector is concentrated primarily in one half of the glacier. The right plot in Figure 6.6 studies these concentrations in each half of the domain using the corresponding L^2-norms. We can distinguish four groups of eigenvectors, highlighted by different colors in Figure 6.6, which we discuss next.

Data-informed eigenvectors

The first group (shown in red in Figure 6.6) contains eigenvectors for which d^i is positive, i.e., the information from the data dominates the information from prior. In the direction of these eigenvectors, the variance in the posterior is significantly reduced due to the observations (recall that the variance is $1/\lambda^i = 1/(r_m^i + r_p^i)$), and hence we say they have been informed by the data.

The eigenvectors in this group are primarily concentrated in the lower half of the glacier where we have observations (see the right plot in Figure 6.6). They are relatively smooth (since r_p^i is not large), and qualitatively resemble the first nine Fourier modes in this region (see plots (a) and (b) in Figure 6.7 for eigenvectors 1 and 3). This last observation is powerful since it provides confidence that features of the MAP point that lie in the span of these first nine Fourier modes are indeed features of the true basal sliding coefficient field.

Shadowed eigenvectors

The next group contains eigenvectors for which the original prior variance was large, and yet the observations provide little information (they are not illuminated by the data, and thus "shadowed"). These eigenvectors are characterized by large ratios r_p^i/r_m^i, and the posterior is easy to characterize in these directions; it is similar to the prior. This group as well as the prior-tail group discussed below are shown in blue in Figure 6.6. Both groups are well characterized by the prior distribution, although for different reasons.

Shadowed eigenvectors generally concentrate in regions where the parameter-to-observable map is insensitive to the parameter. In our problem, the upper part of the glacier is far away from observation points, and the basal sliding coefficient field at a point only has significant influence on the ice velocity in a neighborhood of that point. Thus, the parameter-to-observable map is insensitive to the sliding coefficient in the upper part of the glacier. In Figure 6.6 we can see that indeed most of these eigenvectors are concentrated in the upper part of the glacier, and again resemble Fourier modes in the upper half of the glacier (see Figure 6.7c).

Parameter-to-observable map insensitivity also occurs at the very bottom edge of the glacier. Even though observations are available here, the flow in this region is determined primarily by the glacial boundary, preventing the basal sliding coefficient field from significantly influencing the surface velocity in this region. Exactly one shadowed eigenvector corresponds to this region, shown in Figure 6.7d.

Mixed eigenvectors

The third group contains eigenvectors for which the observations and the prior both have a significant influence. In general it is not clear how this interaction will affect the posterior distribution, and as such it is perhaps too optimistic to make predictions based on this analysis and we defer this discussion to Section 6.6.3. Note that these eigenvectors seem to be generally characterized by a mixture of medium frequency Fourier-like modes on the

upper and lower half of the glacier, which is why we refer to them as "mixed" eigenvectors. One eigenvector from this group is shown in Figure 6.7e.

Prior-tail eigenvectors

The remaining posterior eigenvectors represent directions in which the prior is very certain (i.e., the prior variance $1/r_p^i$ is small), and for which the observations do not provide sufficient evidence to either contradict or reinforce this assertion (i.e., as in the shadowed eigenvectors, the ratio r_p^i/r_m^i is large). In the continuous inverse problem, this final group contains an infinite number of eigenvectors, each behaving very similar to their prior counterparts, and we therefore refer to this as the "prior-tail". One eigenvector from this prior-tail group, which qualitatively resembles a high frequency Fourier mode, is shown in Figure 6.7f.

6.6.3 Marginals in the eigenvector directions

While the analysis of the previous section provides some insight into the posterior distribution, it has two important limitations. First, this analysis is predicated on the assumption that the posterior is completely characterized by its mean and covariance, and as such, any non-Gaussian behavior of the posterior is obscured. Second, the analysis makes use of the posterior covariance approximated at the MAP point, which may not reflect the behavior of the posterior away from this point. In this section, we make use of the insights gleaned from the above analysis, but return our focus to the full posterior

distribution.

In Figure 6.8, we show the one-dimensional marginals and sample variances of the posterior distribution, with respect to eigenvectors of the posterior covariance using colors corresponding to the eigenvector groups discussed in Section 6.6.2. Many features of these marginals are already anticipated: the data-informed eigenvectors (in red) have small variance and are most shifted with respect to the prior distribution. The shadowed eigenvectors (first blue group) have the largest variances and the prior-tail eigenvectors (second blue group) have small variance and are essentially unchanged from the prior. To emphasize the departure of the posterior from the prior, all marginals are plotted with respect to the prior mean. Any shift of the marginal away from zero is due to observations.

Despite the nonlinearity of the parameter-to-observable map, we find that the posterior marginals all appear to be near-Gaussian. Since the noise and prior models are both Gaussian, it is reasonable to expect gaussianity of the data-informed directions in the small-noise limit (the parameter-to-observable map is smooth and thus nearly linear over a narrow range), and also in the directions where the prior is most influential, as the data does not update the prior distribution in these directions. We therefore anticipate that the most non-Gaussian behavior occurs in the mixed eigenvector directions (in green), as these are the directions with the largest variance (so that the parameter-to-observable map can deviate from a linear approximation) that are significantly influenced by the data.

Figure 6.9, depicts one- and two-dimensional marginals of the posterior distribution in selected eigenvector directions together with the Gaussian approximation of the posterior distribution at the MAP point. As in Figure 6.8, these marginals are plotted with respect to the prior mean. In all directions except for the first (the most data-informed eigenvector), we observe that the posterior marginal is close to the Gaussian approximation at the MAP point even in the mixed eigenvector direction (v_{26}). In the direction of the first eigenvector, there is a clear shift in the marginal mean of the posterior distribution and its Gaussian approximation at the MAP point. Nevertheless, the corresponding posterior marginal looks Gaussian. To give a possible explanation for this behavior, consider a two-dimensional pdf with banana-shaped contours for which the MAP point is located along the banana ridge, but the mean is located at the banana's center of mass, in a region that itself may have low probability density. One-dimensional marginals of such a pdf are likely to have a similar discrepancy between the MAP point and the mean. Although with respect to the other eigenvectors, the marginals of the posterior and the Gaussian approximation at the MAP point are close, this does not necessarily imply that the posterior is Gaussian.

6.7 Concluding remarks

We have addressed the problem of constructing efficient MCMC methods for exploring posterior distributions for uncertain parameter fields in infinite-

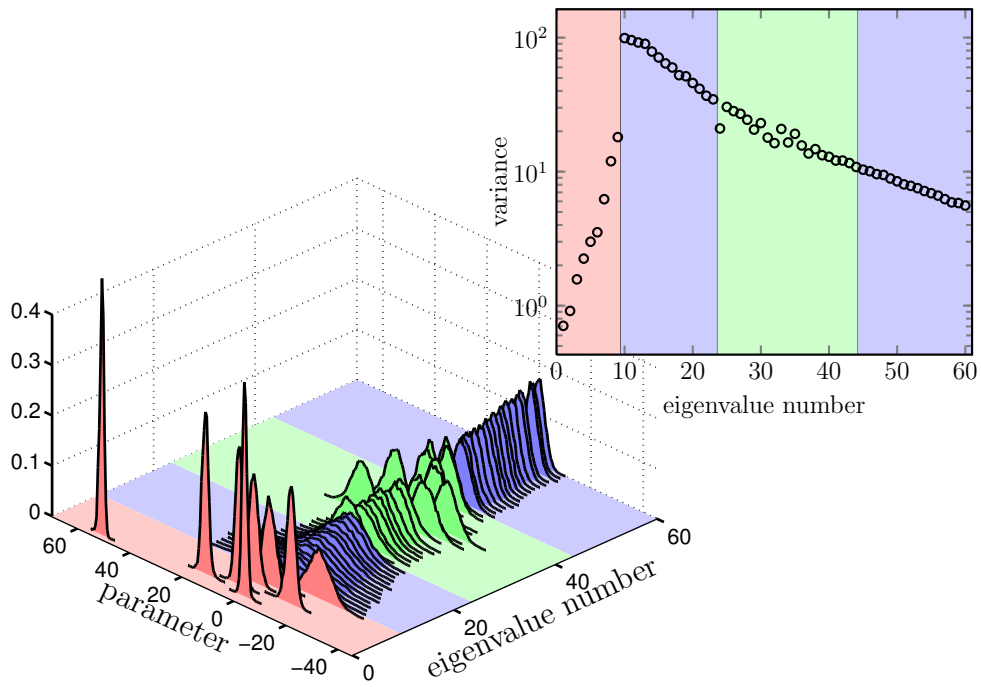

Figure 6.8: Marginals of the posterior distribution with respect to eigenvectors of the covariance (approximated using the Hessian at the MAP point). Shown are the marginals (left) and the corresponding sample variances (right). The eigenvectors are sorted with respect to qualitative features indicated by different background colors as described in the text.

dimensional Bayesian inverse problems governed by expensive forward models. The stochastic Newton MCMC method presented in [143] has been extended in several ways. First, the method is recast in a form that is consistent with the infinite-dimensional setting. In doing so, we have extended the work in [33] to nonlinear inverse problems. Second, the complexity of recomputing the Hessian at each sample point was addressed by investigating a modified stochastic Newton MCMC that reuses the Hessian evaluated at the MAP point.

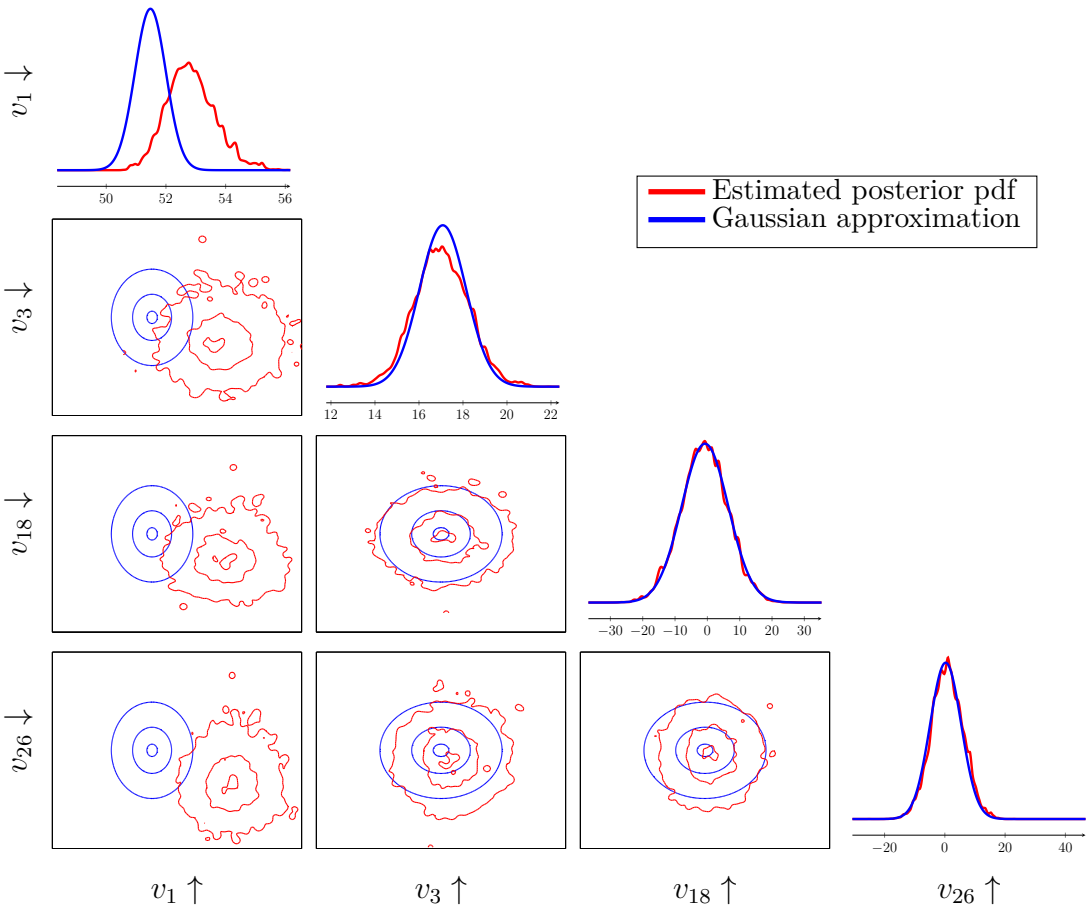

Figure 6.9: One and two-dimensional marginals from the posterior (red) compared with marginals of the Gaussian approximation at the MAP point (blue). The two-dimensional plots show contour lines of the two-dimensional marginals, where the three contours are selected to contain 5%, 50%, and 95% of the density, respectively. The marginals are computed with respect to the eigenvectors v_1, v_3, v_{18} and v_{26}, which are plotted in Figure 6.7a, b, d, and e, respectively. Kernel desity estimation is used to visualize the posterior pdf using the MCMC sample chain.

The modified stochastic Newton MCMC method (with MAP-based Hessian) proposed in this paper is compared with the original stochastic Newton MCMC method (with dynamically changing Hessian) and with an independence sampling method based on a Gaussian proposal at the MAP point, for an ice sheet flow inverse problem governed by a nonlinear Stokes equation. A performance comparison reveals that the proposed stochastic Newton MCMC method with a MAP-based Hessian proposal leads to the best convergence, both in terms of the number of samples as well as in terms of the number of PDE solves.

We also presented visualizations and interpretations of the posterior distribution in high dimensions. We showed point marginals of the posterior to provide intuition about the statistical solution at particular points or regions of the domain. The point marginals confirm the dependence of the variance on the availability of observations. We classified the eigenvectors of the covariance of the Gaussian approximation of the posterior at the MAP into groups depending on the extent to which they are influenced by the observational data versus the prior. This classification can be used to identify and exploit directions in parameter space in which the distribution is Gaussian (for directions that are not informed by the data and hence are dominated by a Gaussian prior) or non-Gaussian (for directions that are informed by the data and hence the nonlinearity of the parameter-to-observable map dominates).

Acknowledgements

We would like to thank Youssef Marzouk for helpful discussions with respect to the interpretation of the posterior distribution.

Chapter 7

Likelihood-informed dimension reduction for nonlinear inverse problems

The content of this chapter is based on an existing publication[1] which is joint work with Tiangang Cui, Youssef Marzouk, Antti Solonen, and Alessio Spantini. Youssef and I collaborated on the germinating ideas for this work, including the ideas of averaging of the prior-preconditioned Hessian over the parameter space and conceptual algorithms for constructing the global likelihood-informed subspace (LIS). Tiangang joined the project later, and all three of us collaborated on convergence metrics, termination criteria, and generation of new parameter states used to construct the LIS. Tiangang ran all of the final numerical experiments for the elliptic problem that appear in this chapter. The GMOS numerical example and results were due to Antti. TC drafted the main body of the paper with some contributions at the later stages from Alessio. All authors contributed significantly with discussion and revisions for the final version of the paper.

[1] T. Cui, J. Martin, Y. M. Marzouk, A. Solonen, and A. Spantini. Likelihood-informed dimension reduction for nonlinear inverse problems. *Inverse Problems*, 30(11):114015, 2014. http://iopscience.iop.org/0266-5611/30/11/114015

Abstract

The intrinsic dimensionality of an inverse problem is affected by prior information, the accuracy and number of observations, and the smoothing properties of the forward operator. From a Bayesian perspective, changes from the prior to the posterior may, in many problems, be confined to a relatively low-dimensional subspace of the parameter space. We present a dimension reduction approach that defines and identifies such a subspace, called the "likelihood-informed subspace" (LIS), by characterizing the relative influences of the prior and the likelihood over the support of the posterior distribution. This identification enables new and more efficient computational methods for Bayesian inference with nonlinear forward models and Gaussian priors. In particular, we approximate the posterior distribution as the product of a lower-dimensional posterior defined on the LIS and the prior distribution marginalized onto the complementary subspace. Markov chain Monte Carlo sampling can then proceed in lower dimensions, with significant gains in computational efficiency. We also introduce a Rao-Blackwellization strategy that de-randomizes Monte Carlo estimates of posterior expectations for additional variance reduction. We demonstrate the efficiency of our methods using two numerical examples: inference of permeability in a groundwater system governed by an elliptic PDE, and an atmospheric remote sensing problem based on Global Ozone Monitoring System (GOMOS) observations.

7.1 Introduction

Inverse problems arise from indirect observations of parameters of interest. The Bayesian approach to inverse problems formalizes the characterization of these parameters through exploration of the *posterior distribution* of parameters conditioned on data [117,189,192]. Computing expectations with respect to the posterior distribution yields not only point estimates of the parameters (e.g., the posterior mean), but a complete description of their uncertainty via the posterior covariance and higher moments, marginal distributions, quantiles, or event probabilities. Uncertainty in parameter-dependent predictions can also be quantified by integrating over the posterior distribution.

The parameter of interest in inverse problems is often a function of space or time, and hence an element of an infinite-dimensional function space [189]. In practice, the parameter field must be discretized, and the resulting inference problem acquires a high but finite dimension. The computation of posterior expectations then proceeds via posterior sampling, most commonly using Markov chain Monte Carlo (MCMC) methods [25,85,140]. The computational cost and efficiency of an MCMC scheme can be strongly affected by the parameter dimension, however. The convergence rates of standard MCMC algorithms usually degrade with parameter dimension [147,168,175,176,177]; one manifestation of this degradation is an increase in the mixing time of the chain, which in turn leads to higher variance in posterior estimates. Some recent MCMC algorithms, formally derived in the infinite-dimensional setting [20,54], do not share this scaling problem. Yet even in this setting, we

will argue that significant variance reduction can be achieved through explicit dimension reduction and through de-randomization of posterior estimates, explained below.

This paper proposes a method for *dimension reduction* in Bayesian inverse problems. We reduce dimension by identifying a subspace of the parameter space that is *likelihood-informed*; this notion will be precisely defined in a relative sense, i.e., relative to the prior. Our focus is on problems with nonlinear forward operators and Gaussian priors, but builds on low-rank approximations [73] and optimality results [186] developed for the linear-Gaussian case. Our dimension reduction strategy will thus reflect the combined impact of prior smoothing, the limited accuracy or number of observations, and the smoothing properties of the forward operator. Identification of the likelihood-informed subspace (LIS) will let us write an approximate posterior distribution wherein the distribution on the complement of this subspace is taken to be independent of the data; in particular, the posterior will be approximated as the product of a low-dimensional posterior on the LIS and the marginalization of the prior onto the complement of the LIS. The key practical benefit of this approximation will be *variance reduction* in the evaluation of posterior expectations. First, Markov chain Monte Carlo sampling can be restricted to coordinates in the likelihood-informed space, enabling greater sampling efficiency—i.e., more independent samples in a given number of MCMC steps or a given computational time. Second, the product form of the approximate posterior will allow sampling in the complement of the likelihood-informed space to be avoided

altogether, thus producing Rao-Blackwellized or analytically conditioned estimates of certain posterior expectations.

Dimension reduction for inverse problems has been previously pursued in several ways. [145] constructs a low dimensional representation of the parameters by using the truncated Karhunen-Lòeve expansion [119, 141] of the prior distribution. A different approach, combining prior and likelihood information via low-rank approximations of the prior-preconditioned Hessian of the log-likelihood, is used in [73] to approximate the posterior covariance in linear inverse problems. In the nonlinear setting, low-rank approximations of the prior-preconditioned Hessian are used to construct proposal distributions in the stochastic Newton MCMC method [143] or to make tractable Gaussian approximations at the posterior mode in [166]—either as a Laplace approximation, as the proposal for an independence MCMC sampler, or as the fixed preconditioner for a stochastic Newton proposal. We note that these schemes bound the tradeoff between evaluating Hessian information once (at the posterior mode) or with every sample (in local proposals). In all cases, however, MCMC sampling proceeds in the full-dimensional space.

The dimension reduction approach explored in this paper, by contrast, confines sampling to a lower-dimensional space. We extend the posterior approximation proposed in [186] to the nonlinear setting by making essentially a low-rank approximation of the *posterior expectation* of the prior-preconditioned Hessian, from which we derive a projection operator. This projection operator then yields the product-form posterior approximation dis-

cussed above, which enables variance reduction through lower-dimensional MCMC sampling and Rao-Blackwellization of posterior estimates.

We note that our dimension reduction approach does not depend on the use of any specific MCMC algorithm, or even on the use of MCMC. The low-dimensional posterior defined on coordinates of the LIS is amenable to a range of posterior exploration or integration approaches. We also note that the present analysis enables the construction of dimension-independent analogues of existing MCMC algorithms with essentially no modification. This is possible because in inverse problems with formally discretization-invariant posteriors—i.e., problems where the forward model converges under mesh refinement and the prior distribution satisfies certain regularity conditions [125, 189]—the LIS can also be discretization invariant. We will demonstrate these discretization-invariance properties numerically.

The rest of this paper is organized as follows. In Section 2, we briefly review the Bayesian formulation for inverse problems. In Section 3, we introduce the likelihood-informed dimension reduction technique, and present the posterior approximation and reduced-variance Monte Carlo estimators based on the LIS. We also present an algorithm for constructing the likelihood-informed subspace. In Section 4, we use an elliptic PDE inverse problem to demonstrate the accuracy and computational efficiency of our posterior estimates and to explore various properties of the LIS, including its dependence on the data and its discretization invariance. In Section 5, we apply our variance reduction technique to an atmospheric remote sensing problem. Section 6 offers

concluding remarks.

7.2 Bayesian formulation for inverse problems

This section provides a brief overview of the Bayesian framework for the inverse problems as introduced in [117,189,192]. Consider the inverse problem of estimating parameters x from data y, where

$$y = G(x) + e\,. \tag{7.1}$$

Here e is a random variable representing noise and/or model error, which appears additively, and G is a known mapping from the parameters to the observables. In a Bayesian setting, we model the parameters x as a random variable and, for simplicity, assume that the range of this random variable is a finite dimensional space $\mathbb{X} \subseteq \mathbb{R}^n$. Then the parameter of interest is characterized by its posterior distribution conditioned on a realization of the data, $y \in \mathbb{Y} \subseteq \mathbb{R}^d$:

$$\pi(x|y) \propto \pi(y|x)\pi_0(x). \tag{7.2}$$

We assume that all distributions have densities with respect to Lebesgue measure. The posterior probability density function above is the product of two terms: the prior density $\pi_0(x)$, which models knowledge of the parameters before the data are observed, and the likelihood function $\pi(y|x)$, which describes the probability distribution of y for any value of x.

We assume that the prior distribution is a multivariate Gaussian $\mathcal{N}(\mu_{\text{pr}}, \Gamma_{\text{pr}})$, where the covariance matrix Γ_{pr} can be also defined by its inverse, Γ_{pr}^{-1}, com-

monly referred to as the precision matrix. We model the additive noise with a zero mean Gaussian distribution, i.e., $e \sim \mathcal{N}(0, \Gamma_{\text{obs}})$. This lets us define the data-misfit function

$$\eta(x) = \frac{1}{2} \left\| \Gamma_{\text{obs}}^{-\frac{1}{2}} \left(G(x) - y \right) \right\|^2, \tag{7.3}$$

such that the likelihood function is proportional to $\exp\left(-\eta(x)\right)$.

7.3 Methodology

7.3.1 Optimal dimension reduction for linear inverse problems

Consider a linear forward model, $G(x) = Gx$, with a Gaussian likelihood and a Gaussian prior as defined in Section 7.2. The resulting posterior is also Gaussian, $\pi(x|y) = \mathcal{N}(\mu_{\text{pos}}, \Gamma_{\text{pos}})$, with mean and covariance given by

$$\mu_{\text{pos}} = \Gamma_{\text{pos}} \left(\Gamma_{\text{pr}}^{-1} \mu_{\text{pr}} + G^\top \Gamma_{\text{obs}}^{-1} y \right) \quad \text{and} \quad \Gamma_{\text{pos}} = \left(H + \Gamma_{\text{pr}}^{-1} \right)^{-1}, \tag{7.4}$$

where $H = G^\top \Gamma_{\text{obs}}^{-1} G$ is the Hessian of the data-misfit function (7.3). Without loss of generality we can assume zero prior mean and a positive definite prior covariance matrix.

Now consider approximations to the posterior distribution of the form

$$\tilde{\pi}(x|y) \propto \pi\left(y|P_r x\right) \pi_0(x), \tag{7.5}$$

where $P_r = P_r^2$ is a rank-r projector and $\pi\left(y|P_r x\right)$ is an approximation to the original likelihood function $\pi\left(y|x\right)$. Approximations of this form can be computationally advantageous when operations involving the prior (e.g., evaluations or sampling) are less expensive than those involving the likelihood.

As described in [186], they are also the natural form with which to approximate a Bayesian update, particularly in the inverse problem setting with high-dimensional x. In the deterministic case, inverse problems are ill-posed; the data cannot inform all directions in the parameter space. Equivalently, the spectrum of H is compact or decays quickly. Thus one should be able to explicitly project the argument of the likelihood function onto a low-dimensional space without losing much information in the process. The posterior covariance remains full rank, but the update from prior covariance to posterior covariance will be low rank. The challenge, of course, is to find the best projector P_r for any given r. The answer will involve balancing the influence of the prior and the likelihood. In the following theorem, we introduce the optimal projector and characterize its approximation properties.

Theorem 4. *Let $\Gamma_{\text{pr}} = LL^\top$ be a symmetric decomposition of the prior covariance matrix and let (λ_i, v_i) be the eigenvalue-eigenvector pairs of the prior-preconditioned Hessian $\left(L^\top HL\right)$ such that $\lambda_i \geq \lambda_{i+1}$. Define the directions $u_i = Lv_i$ and $w_i = L^{-\top}v_i$ together with the matrices $U_r = [u_1, \ldots, u_r]$ and $W_r = [w_1, \ldots, w_r]$. Then, the projector P_r given by:*

$$P_r = U_r W_r^\top,$$

yields an approximate posterior density of the form $\tilde{\pi}(x|y) = \mathcal{N}\left(\mu_{\text{pos}}^{(r)}, \Gamma_{\text{pos}}^{(r)}\right)$ and is optimal in the following sense:

1. *$\Gamma_{\text{pos}}^{(r)}$ minimizes the Förstner distance [74] from the exact posterior covariance over the class of positive definite matrices that can be written*

as rank r negative semidefinite updates of the prior covariance.

2. *$\mu_{\text{pos}}^{(r)} = A^* y$ minimizes the Bayes risk $\mathbb{E}_{x,y}\left[\|\mu(y) - x\|_{\Gamma_{\text{pos}}^{-1}}^2\right]$ over the class of all linear transformations of the data $\mu(y) = Ay$ with $\operatorname{rank}(A) \leq r$.*

Proof. We refer the reader to [186] for a proof and detailed discussion. □

The vectors $(u_1, \ldots, u_r)$ span the range of the optimal projector; we call this range the *likelihood-informed subspace* of the linear inverse problem. Note that the (u_i) are eigenvectors of the pencil $(H, \Gamma_{\text{pr}}^{-1})$. Hence, the jth basis vector u_j maximizes the Rayleigh quotient

$$\mathcal{R}(u) = \frac{\langle u, Hu \rangle}{\langle u, \Gamma_{\text{pr}}^{-1} u \rangle} \tag{7.6}$$

over the subspace $\mathbb{X} \setminus \operatorname{span}\{u_1, \ldots, u_{j-1}\}$. This Rayleigh quotient helps interpret the (u_i) as directions where the data are most "informative" relative to the prior. For example, consider a direction $w \in \mathbb{X}$ representing a rough mode in the parameter space. If the prior is smoothing, then the denominator of (7.6) will be large; also, if the forward model output is relatively insensitive to variation in the w direction, the numerator of (7.6) will be small. Thus the Rayleigh quotient will be small and w is not particularly data-informed relative to the prior. Conversely, if w is smooth then the prior variance in this direction may be large and the likelihood may be relatively constraining; this direction is then data-informed. Of course, there are countless intermediate cases, but in general, directions for which (7.6) are large will lie in the range of U_r.

Note also that U_r diagonalizes both H and Γ_{pr}^{-1}. We are particularly interested in the latter property: the modes (u_i) are orthogonal (and can be chosen orthonormal) with respect to the inner product induced by the prior precision matrix. This property will be preserved later in the nonlinear case, and will be important to our posterior sampling schemes.

For nonlinear inverse problems, we seek an approximation of the posterior distribution in the same form as (7.5). In particular, the range of the projector will be determined by blending together local likelihood-informed subspaces from regions of high posterior probability. The construction of the approximation will be detailed in the following section.

7.3.2 LIS construction for nonlinear inverse problems

When the forward model is nonlinear, the Hessian of the data-misfit function varies over the parameter space, and thus the likelihood-informed directions are embedded in some nonlinear manifold. We aim to construct a global linear subspace to capture the majority of this nonlinear likelihood-informed manifold.

Let the forward model $G(x)$ be first-order differentiable. The linearization of the forward model at a given parameter value x, $J(x) \equiv \nabla G(x)$ where $J(x) \in \mathbb{R}^{d\times n}$, provides the local sensitivity of the parameter-to-observable map. Inspired by the dimension reduction approach for the linear inverse problem, we use the linearized forward model $J(x)$ to construct the Gauss-Newton approximation of the Hessian of the data-misfit function, $H(x) =$

$J(x)^\top \Gamma_{\text{obs}}^{-1} J(x)$. Now consider a local version of the Rayleigh quotient (7.6),

$$\mathcal{R}(u;x) := \frac{\langle u, H(x)u\rangle}{\langle u, \Gamma_{\text{pr}}^{-1}u\rangle}.$$

Introducing the change of variable $v = L^{-1}u$, we can equivalently use

$$\widetilde{\mathcal{R}}(v;x) := \frac{\langle v, (L^\top H(x)L)v\rangle}{\langle v, v\rangle} = \mathcal{R}(Lv;x), \tag{7.7}$$

to quantify the local impact of the likelihood relative to the prior. As in the linear problem, this suggests the following procedure for computing a local LIS given some truncation threshold τ_{loc}:

Problem 1 (Construction of the local likelihood-informed subspace)**.** *Given the Gauss-Newton Hessian of the data misfit function $H(x)$ at a given x, find the eigendecompostion of the prior-preconditioned Gauss-Newton Hessian (ppGNH)*

$$L^\top H(x) L v_i = \lambda_i v_i. \tag{7.8}$$

Given a truncation threshold $\tau_{loc} > 0$, the local LIS is spanned by $U_l = [u_1, \ldots, u_l]$, where $u_i = Lv_i$ corresponds to the l leading eigenvalues such that $\lambda_1 \geq \lambda_2 \geq \ldots \geq \lambda_l \geq \tau_{loc}$.

For a direction u with $\mathcal{R}(u;x) = 1$, the local impact of the likelihood and the prior are balanced. Thus, to retain a comprehensive set of likelihood-informed directions, we typically choose a truncation threshold τ_{loc} less than 1.

To extend the pointwise criterion (7.7) into a global criterion for likelihood-informed directions, we consider the expectation of the Rayleigh quotient over the posterior

$$\mathbb{E}_\pi\left[\mathcal{R}(u;x)\right] = \mathbb{E}_\pi\left[\widetilde{\mathcal{R}}(v;x)\right] = \frac{\langle v, Sv\rangle}{\langle v, v\rangle},$$

where S is the expected ppGNH over the posterior,

$$S = \int_{\mathbb{X}} L^\top H(x) L \, \pi(dx|y). \tag{7.9}$$

Then we can naturally construct the global LIS through the eigendecomposition of S as in the linear case. We consider approximating S using the Monte Carlo estimator

$$\widehat{S}_n = \frac{1}{n}\sum_{k=1}^{n} L^\top H(x^{(k)}) L,$$

where $x^{(k)} \sim \pi(x|y)$, $k = 1 \dots n$, are posterior samples. Since the local Hessian $H(x^{(k)})$ is usually not explicitly available and is not feasible to store for large-scale problems, we use its prior-preconditioned low-rank approximation as defined in Problem 1. Thus the global LIS can be constructed by the following procedure:

Problem 2 (Construction of global likelihood-informed subspace)**.** *Suppose we have a set of posterior samples $\mathcal{X} = \{x^{(k)}\}$, $k = 1 \dots m$, where for each sample $x^{(k)}$, the ppGNH is approximated by the truncated low rank eigendecomposition*

$$L^\top H(x^{(k)}) L \approx \sum_{i=1}^{l(k)} \lambda_i^{(k)} v_i^{(k)} v_i^{(k)\top},$$

by solving Problem 1. We have $\lambda_i^{(k)} \geq \tau_{loc}$ for all $k = 1 \dots m$ and all $i = 1 \dots l(k)$. To construct the global LIS, we consider the eigendecompostion of

the Monte Carlo estimator of the expected Hessian in (7.9), which takes the form

$$\left(\frac{1}{m}\sum_{k=1}^{m}\sum_{i=1}^{l(k)}\lambda_i^{(k)}v_i^{(k)}v_i^{(k)\top}\right)\psi_j=\gamma_j\psi_j. \tag{7.10}$$

The global LIS has the non-orthogonal basis $\Phi_r = L\Psi_r$*, where the eigenvectors* $\Psi_r = [\psi_1, \ldots, \psi_r]$ *correspond to the* r *leading eigenvalues of (7.10),* $\gamma_1 \geq \ldots \geq \gamma_r \geq \tau_g$*, for some truncation threshold* $\tau_g > 0$*. Here we choose* τ_g *to be equal to the threshold* τ_{loc} *in Problem 1.*

In many applications we can only access the Gauss-Newton Hessian by computing its action on vectors, which involves one forward model evaluation and one adjoint model evaluation. In such a case, the ppGNH can be approximated by finding the eigendecomposition (7.8) using either Krylov subspace algorithms [89] or randomized algorithms [98, 134].

The number of samples required to construct the global LIS depends on the degree to which $H(x)$ (or its dominant eigenspace) varies over the posterior. In Section 7.3.5, we present an adaptive construction procedure that automatically explores the directional landscape of the likelihood.

7.3.3 Posterior approximation

By projecting the likelihood function onto the global likelihood-informed subspace (LIS), we obtain the approximate posterior

$$\tilde{\pi}(x|y) \propto \pi\left(y|\Pi_r x\right)\pi_0(x), \tag{7.11}$$

where Π_r is a projector onto the global LIS. This projector is self-adjoint with respect to the inner product induced by the prior precision matrix, and leads to a natural decomposition of the parameter space as $\mathbb{X} = \mathbb{X}_r \oplus \mathbb{X}_\perp$, where $\mathbb{X}_r = \text{range}(\Pi_r)$ is the LIS and $\mathbb{X}_\perp = \text{range}(I - \Pi_r)$ is the complement subspace (CS). This choice leads to a factorization of the prior distribution into the product of two distributions, one defined on the low-dimensional LIS and the other on the CS. This factorization is the key to our dimension reduction technique.

Definition 2. *We define the projectors Π_r and $I - \Pi_r$, and a corresponding parameter decomposition, as follows:*

(a) Suppose the LIS basis computed in Problem 2 is $\Phi_r = L\Psi_r$, where Ψ_r is orthonormal. Define the matrix $\Xi_r = L^{-\top}\Psi_r$ such that $\Xi_r^\top \Phi_r = I_r$. The projector Π_r has the form

$$\Pi_r = \Phi_r \Xi_r^\top.$$

Choose $\Psi_\perp$ such that $[\Psi_r \ \Psi_\perp]$ forms a complete orthonormal system in $\mathbb{R}^n$. Then the projector $I - \Pi_r$ can be written as

$$I - \Pi_r = \Phi_\perp \Xi_\perp^\top,$$

where $\Phi_\perp = L\Psi_\perp$ and $\Xi_\perp = L^{-\top}\Psi_\perp$.

(b) Naturally, the parameter x can be decomposed as

$$x = \Pi_r x + (I - \Pi_r)x,$$

where each projection can be represented as the linear combination of the corresponding basis vectors. Consider the "LIS parameter" x_r and the "CS parameter" $x_\perp$, which are the weights associated with the LIS basis Φ_r and CS basis $\Phi_\perp$, respectively. Then we can define the following pair of linear transformations between the parameter x and $(x_r, x_\perp)$:

$$x = [\Phi_r \ \Phi_\perp] \begin{bmatrix} x_r \\ x_\perp \end{bmatrix}, \quad \begin{bmatrix} x_r \\ x_\perp \end{bmatrix} = [\Xi_r \ \Xi_\perp]^\top x. \tag{7.12}$$

Figure 7.1 illustrates the transformations between the parameter projected onto the LIS, $\Pi_r x$, and the LIS parameter x_r. The same relation holds for the transformations between $(I - \Pi_r)x$ and the CS parameter $x_\perp$. And as Definition 2 makes clear, Π_r is an oblique projector.

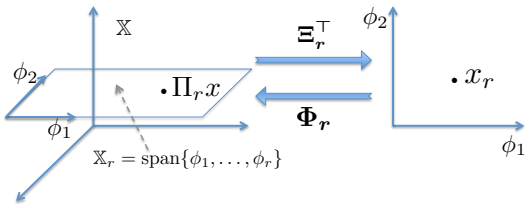

Figure 7.1: Illustration of the transformation between the parameter projected onto the LIS, $\Pi_r x$, and the LIS parameter x_r.

Lemma 5. *Suppose we have $x = \Phi_r x_r + \Phi_\perp x_\perp$ as defined in Definition 2(b). Then the prior distribution can be decomposed as*

$$\pi_0(x) = \pi_r(x_r)\pi_\perp(x_\perp),$$

where $\pi_r(x_r) = \mathcal{N}(\Xi_r^\top \mu_{\text{pr}}, I_r)$ and $\pi_\perp(x_\perp) = \mathcal{N}(\Xi_\perp^\top \mu_{\text{pr}}, I_\perp)$.

Following Definition 2(b) and Lemma 5, the approximate posterior distribution (7.11) can be reformulated as

$$\begin{aligned}\tilde{\pi}(x|y) &\propto \pi\left(y|\Pi_r x\right)\pi_r(x_r)\pi_\perp(x_\perp)\\ &= \pi\left(y|\Phi_r x_r\right)\pi_r(x_r)\pi_\perp(x_\perp).\end{aligned}$$

Applying the linear transformation from x to $(x_r, x_\perp)$ as defined in Equation (7.12), we can rewrite the approximate posterior for the parameters $(x_r, x_\perp)$ as

$$\tilde{\pi}(x_r, x_\perp|y) \propto \tilde{\pi}(x|y) \propto \tilde{\pi}(x_r|y)\pi_\perp(x_\perp), \tag{7.13}$$

which is the product of the reduced posterior

$$\tilde{\pi}(x_r|y) \propto \pi\left(y|\Phi_r x_r\right)\pi_r(x_r), \tag{7.14}$$

and the complement prior $\pi_\perp(x_\perp)$. To compute a Monte Carlo estimate of the expectation of a function over the approximate posterior distribution (7.13), we only need to sample the reduced posterior $\tilde{\pi}\left(x_r|y\right)$, since properties of the Gaussian complement prior $\pi_\perp(x_\perp)$ are known analytically.

One can also combine MCMC samples from the reduced posterior $\tilde{\pi}\left(x_r|y\right)$ with independent samples from the complement prior $\pi_\perp(x_\perp)$ to provide samples that are approximately drawn from the full posterior $\pi(x|y)$. By correcting these samples via importance weights or a Metropolis scheme, one would then obtain a sampling algorithm for the original full-space posterior. This idea is not pursued further here, and in the rest of this work we will emphasize the analytical properties of the complement prior $\pi_\perp(x_\perp)$, using them to reduce the variance of Monte Carlo estimators.

7.3.4 Reduced-variance estimators

Suppose we have a function $h(x)$ for which the conditional expectation over the approximate posterior (7.11)

$$\mathbb{E}_{\tilde{\pi}}\left[h(x)|x_r\right] = \int_{\mathbb{X}_\perp} h(\Phi_r x_r + \Phi_\perp x_\perp)\, \pi_0(x_\perp)\, dx_\perp, \tag{7.15}$$

can be calculated either analytically or through some high-precision numerical quadrature scheme. Then variance reduction can be achieved as follows:

1. **Subspace MCMC.** Use MCMC in the LIS to simulate a "subspace Markov chain" with target distribution $\tilde{\pi}(x_r|y)$ (7.13). Any number of MCMC algorithms developed in the literature can be applied off-the-shelf, e.g., adaptive MCMC [7, 11, 96, 178, 179], the stochastic Newton algorithm of [143], or the Riemannian manifold algorithms of [86]. Since the dimension of the LIS can be quite small relative to the original parameter space, the subspace MCMC approach can yield lower sample correlations (better mixing) than applying any of these MCMC algorithms directly to the full posterior (7.2).

2. **Rao-Blackwellization.** We approximate $\mathbb{E}_\pi[h] = \int_{\mathbb{X}} h(x)\pi(dx|y)$ by the expectation of the function $h(x)$ over the approximate posterior $\tilde{\pi}(x|y)$, i.e., $\mathbb{E}_{\tilde{\pi}}[h] = \int_{\mathbb{X}} h(x)\tilde{\pi}(dx|y)$. Given a set of subspace MCMC samples $\{x_r^{(1)}, \ldots, x_r^{(N)}\}$ where $x_r^{(k)} \sim \tilde{\pi}_r(x_r|y)$, a Monte Carlo estimator of $\mathbb{E}_{\tilde{\pi}}[h]$ is given by

$$\tilde{Q}_N = \frac{1}{N}\sum_{k=1}^{N} \mathbb{E}_{\tilde{\pi}}\left[h(x)|x_r^{(k)}\right]. \tag{7.16}$$

As an application of the Rao-Blackwellization principle (see [44] and references therein), the estimator (7.16) has a lower variance than the standard estimator

$$Q_N = \frac{1}{N}\sum_{k=1}^{N} h(x^{(k)}), \tag{7.17}$$

where $x^{(k)} \sim \pi(x|y)$.

This procedure mitigates many of the difficulties of posterior exploration in high dimensions, provided that the prior-to-posterior update is reasonably low rank. Variance reduction is achieved not only by increasing the effective sample size per MCMC iteration (via subspace MCMC), but also by reducing the variance of Monte Carlo estimators using Rao-Blackwellization. In effect, we argue that the Gaussian CS can be explored *separately* and via a calculation (7.15) that does not involve sampling.

Note that while this procedure necessarily reduces the variance of a Monte Carlo estimator, it introduces bias since we replace the expectation over the full posterior $\mathbb{E}_\pi[h]$ with an expectation over the approximate posterior $\mathbb{E}_{\tilde{\pi}}[h]$. Thus this variance reduction is particularly useful in situations where the variance of the estimator (7.17) derived from full-space MCMC samples is large compared with the bias, which is often the case for high-dimensional inverse problems.

Beyond variance reduction, subspace MCMC offers several additional computational advantages over MCMC methods applied to the full posterior directly: (i) The storage requirement for saving subspace MCMC samples is

much lower than that of an MCMC scheme that samples the full posterior. (ii) For MCMC methods where the proposal distribution involves operations with square root of the prior covariance matrix (e.g., the stochastic Newton [143] and preconditioned Crank-Nicolson [20,54] techniques) the computational cost of handling the full prior covariance can be much higher than the computational cost of handling the reduced prior $\pi_r(x_r)$, which has identity covariance.

The Monte Carlo estimator (7.16) can be further simplified if the function of interest $h(x)$ can be expressed as either the product or the sum of two separate functions, $h_r(x_r)$ and $h_\perp(x_\perp)$, defined on the LIS and CS, respectively. In the multiplicative case $h(x) = h_r(x_r)h_\perp(x_\perp)$, the conditional expectation (7.15) can be written as

$$\mathbb{E}_{\tilde{\pi}}\left[h(x)|x_r\right] = h_r(x_r)\int_{\mathbb{X}_\perp} h_\perp(x_\perp)\pi_0(dx_\perp).$$

In the additive case $h(x) = h_r(x_r) + h_\perp(x_\perp)$, it can be written as

$$\mathbb{E}_{\tilde{\pi}}\left[h(x)|x_r\right] = h_r(x_r) + \int_{\mathbb{X}_\perp} h_\perp(x_\perp)\pi_0(dx_\perp).$$

Thus the expectation $\mathbb{E}_{\tilde{\pi}}[h]$ can be decomposed either into the product (in the multiplicative case) or the sum (in the additive case) of the pair of expectations

$$\mathbb{E}_{\tilde{\pi}}[h_r] = \int_{\mathbb{X}_r} h_r(x_r)\pi(dx_r|y), \tag{7.18}$$

$$\mathbb{E}_{\tilde{\pi}}[h_\perp] = \int_{\mathbb{X}_\perp} h_\perp(x_\perp)\pi_0(dx_\perp), \tag{7.19}$$

which are associated with the LIS and CS, respectively. The expectation in (7.18) can be computed by the subspace MCMC methods described above,

whereas the expectation in (7.19) is computed analytically or through high-order numerical integration.

Now we give two particularly useful examples of the analytical treatment of the complement space.

Example 1 (Reduced variance estimator of the posterior mean)**.** *Suppose we have obtained the empirical posterior mean $\tilde{\mu}_r$ of the reduced parameter x_r using subspace MCMC. The resulting reduced-variance estimator of the posterior mean is*

$$\mathbb{E}_{\tilde{\pi}}[x] = \Phi_r \tilde{\mu}_r + \Pi_\perp \mu_{\text{pr}} = \Phi_r \tilde{\mu}_r + (I - \Pi_r)\mu_{\text{pr}}.$$

Example 2 (Reduced variance estimator of the posterior covariance)**.** *Suppose we have the empirical posterior covariance $\tilde{\Gamma}_r$ of the reduced parameter x_r, estimated using subspace MCMC. The resulting reduced-variance estimator of the posterior covariance is*

$$\begin{aligned} \mathbb{C}\text{ov}_{\tilde{\pi}}[x] &= \Phi_r \tilde{\Gamma}_r \Phi_r^\top + \Pi_\perp \Gamma_{\text{pr}} \Pi_\perp^\top \\ &= \Gamma_{\text{pr}} + \Phi_r \left(\tilde{\Gamma}_r - I_r \right) \Phi_r^\top. \end{aligned}$$

7.3.5 Algorithms for the LIS

Constructing the global LIS requires a set of posterior samples. Since the computational cost of solving Problem 1 for any sample is much greater than the cost of evaluating the forward model, we wish to limit the number of samples used in Problem 2 while ensuring that we adequately capture the

posterior variation of the ppGNH. Thus we choose samples using the following adaptive procedure.

Algorithm 1 (Global LIS construction using subspace MCMC)**.** *First, compute the posterior mode $x_{map} \in \mathbb{X}$. Set the initial sample set for Problem 2 to $\mathcal{X}^{(1)} = \{x_{map}\}$. Solve Problem 2 to find $\Psi_r^{(1)}$, the initial LIS basis $\Phi_r^{(1)}$, and its left-inverse $\Xi_r^{(1)}$.*[2] *Initialize a subspace Markov chain with initial state $\Xi_r^{(1)\top} x_{map}$, which is the posterior mode projected onto the LIS. At any subsequent step $k \geq 1$, the following procedure is used to adaptively enrich the LIS:*

1. **Subchain simulation.** *Simulate the $r(k)$-dimensional subspace MCMC chain for L iterations, so that the last state of this chain, denoted by θ, is uncorrelated with its initial state. Then θ transformed back to the original parameter space, $(\Phi_r^{(k)}\theta)$, is used as the next sample point. Enrich the sample set to $\mathcal{X}^{(k+1)} = \mathcal{X}^{(k)} \cup \{\Phi_r^{(k)}\theta\}$.*

2. **LIS construction.** *Solve Problem 2 with the sample set $\mathcal{X}^{(k+1)}$. Then update the LIS basis to $\Phi_r^{(k+1)}$ and $\Xi_r^{(k+1)}$. Set the initial state of the next subspace MCMC chain to $\Xi_r^{(k+1)\top}\Phi_r^{(k)}\theta$.*

3. **Convergence checking.** *Terminate the adaptation if a pre-specified maximum allowable number of Hessian evaluations is exceeded, or if the*

[2] The dimension of the global LIS can vary at each iteration. Let $r(k)$ denote the dimension of the global LIS at iteration k. To be precise, we should then write $\Phi_{r(k)}^{(k)}$ and $\Xi_{r(k)}^{(k)}$, but for brevity we will simplify notation to $\Phi_r^{(k)}$ and $\Xi_r^{(k)}$ when possible.

weighted subspace distance in Definition 3 falls below a certain threshold. Otherwise, set $k \leftarrow k+1$ and return to Step (1).

The convergence criterion in step (3) is based on an incremental distance between likelihood-informed subspaces. The distance penalizes changes in the dominant directions (those with large eigenvalues γ) more heavily than changes in the less important directions (those with smaller γ).

Definition 3 (Weighted subspace distance)**.** *At iteration k, define the basis/weights pair $\mathcal{Y}^{(k)} = \{\Psi_r^{(k)}, D^{(k)}\}$, where $\Psi_r^{(k)}$ is the orthonormal LIS basis from Problem 2 and $D_{ij}^{(k)} = \delta_{ij}(\hat{\gamma}_i^{(k)})^{\frac{1}{4}}$ is a diagonal matrix consisting of normalized weights*

$$\hat{\gamma}_i^{(k)} = \frac{\gamma_i^{(k)}}{\sum_{j=1}^{r(k)} \gamma_j^{(k)}}, \quad j = 1 \dots r(k),$$

computed from the eigenvalues $\{\gamma_1^{(k)}, \dots, \gamma_{r(k)}^{(k)}\}$ of Problem 2. For two adjacent steps k and $k+1$, we compute the weighted subspace distance of [132], which has the form

$$\mathcal{D}\left(\mathcal{Y}^{(k)}, \mathcal{Y}^{(k+1)}\right) = \sqrt{1 - \left\| \left(\Psi_{r(k)}^{(k)} D^{(k)}\right)^\top \left(\Psi_{r(k+1)}^{(k+1)} D^{(k+1)}\right) \right\|_F^2}. \tag{7.20}$$

Note that in Step (i) of Algorithm 8, we construct the global LIS by always sampling in an adaptively enriched subspace. This offers computational benefits, since the MCMC exploration is always confined to a lower dimensional space. However, a potential problem with this approach is that it might ignore some directions that are also data-informed. A more conservative approach would be to introduce a *conditional update* at the end of each subchain

simulation: perform Metropolized independence sampling in the current CS using the complement prior as proposal. This would enable the subchain to explore the full posterior, but would result in higher-dimensional sampling when constructing the LIS. In our numerical examples, described below, no conditional updates were required for good performance; constructing the LIS using samples from the full posterior and using the subspace approach gave essentially the same results. Of course, one could also simply employ a standard MCMC algorithm to sample the full posterior, and then construct the LIS using the resulting posterior samples. However, the efficiency of the MCMC algorithm in this case will be affected by the dimensionality of the problem.

7.4 Example 1: Elliptic PDE

Our first example is an elliptic PDE inverse problem used to demonstrate (i) construction of the LIS and the impact of mesh refinement; (ii) the application of low-rank posterior mean and variance estimators; and (iii) changes in the LIS with varying amounts of observational data.

7.4.1 Problem setup

Consider the problem domain $\Omega = [0, 3] \times [0, 1]$, with boundary $\partial\Omega$. We denote the spatial coordinate by $s \in \Omega$. Consider the permeability field $\kappa(s)$, the pressure field $p(s)$, and sink/source terms $f(s)$. The pressure field for a given permeability and source/sink configuration is governed by the Poisson

equation

$$\begin{cases} -\nabla \cdot (\kappa(s)\nabla p(s)) &= f(s), & s \in \Omega \\ \langle \kappa(s)\nabla p(s), \vec{n}(s) \rangle &= 0, & s \in \partial\Omega \end{cases} \tag{7.21}$$

where $\vec{n}(s)$ is the outward normal vector on the boundary. To make a well-posed boundary value problem, a further boundary condition

$$\int_{\partial\Omega} p(s)ds = 0, \tag{7.22}$$

is imposed. The source/sink term $f(s)$ is defined by the superposition of four weighted Gaussian plumes with standard deviation (i.e., spatial width) 0.05, centered at four corners $[0,0]$, $[3,0]$, $[3,1]$ and $[0,1]$, with weights $\{1,2,3,-6\}$. The system of equations (7.21) is solved by the finite element method with 120×40 bilinear elements.

The discretized permeability field κ is endowed with a log-normal prior distribution, i.e.,

$$\kappa = \exp(x), \text{ and } x \sim \mathcal{N}(0, \Gamma_{\text{pr}}), \tag{7.23}$$

where the covariance matrix Γ_{pr} is defined through an anisotropic exponential covariance kernel

$$\mathbb{C}\text{ov}(x(s), x(s')) = \sigma_u^2 \exp\left(-\frac{\left((s-s')^\top \Sigma^{-1}(s-s')\right)^{\frac{1}{2}}}{s_0} \right), \tag{7.24}$$

for $s, s' \in \Omega$. In this example, we set the anisotropic correlation tensor to

$$\Sigma = \begin{bmatrix} 0.55 & -0.45 \\ -0.45 & 0.55 \end{bmatrix},$$

the prior standard deviation to $\sigma_u = 1.15$, and the correlation length to $s_0 = 0.18$. The "true" permeability field is a realization from the prior distribution.

The true permeability field, the sources/sinks, the simulated pressure field, and the synthetic data are shown in Figure 7.2.

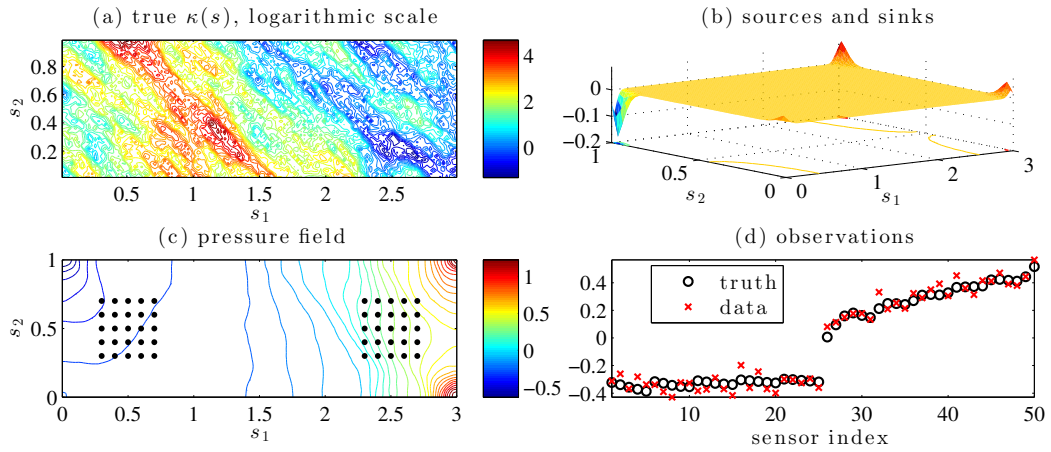

Figure 7.2: Setup of the elliptic inversion example. (a) "True" permeability field. (b) Sources and sinks. (c) Pressure field resulting from the true permeability field, with measurement sensors indicated by black circles. (d) Data y; circles represent the noise-free pressure at each sensor, while crosses represent the pressure observations corrupted with measurement noise.

Partial observations of the pressure field are collected at 50 measurement sensors as shown by the black dots in Figure 7.2(c). The observation operator M is simply the corresponding "mask" operation. This yields observed data $y \in \mathbb{R}^{50}$ as

$$y = Mp(s) + e,$$

with additive error $e \sim \mathcal{N}(0, \sigma^2 I_{50})$. The standard deviation σ of the measurement noise is prescribed so that the observations have signal-to-noise ratio 10, where the signal-to-noise ratio is defined as $\max_s\{p(s)\}/\sigma$. The noisy data are shown in Figure 7.2(d).

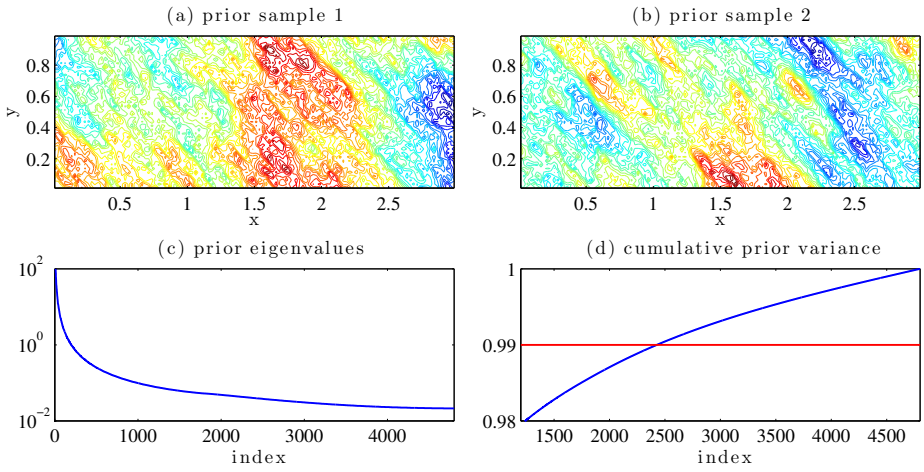

Figure 7.3: Prior samples and eigenspectrum of the prior covariance. (a) and (b): Two samples drawn from the prior. (c) Prior covariance spectrum, eigenvalues versus index number. (d) Cumulative energy (integrated prior variance) over a subset of the eigenspectrum, shown in blue; the red line represents the 99% energy truncation threshold.

Figure 7.3 shows two draws from the prior distribution, the eigenspectrum of the prior covariance, and the cumulative prior variance integrated over Ω (i.e., the running sum of the prior covariance eigenvalues). In order to keep 99% percent of the energy in the prior, 2427 eigenmodes are required. Because of this slow decay of the prior covariance spectrum, *a priori* dimension reduction based on a truncated eigendecomposition of the prior covariance (as described in [145]) would be very inefficient for this problem. Information carried in high-frequency eigenfunctions cannot be captured unless an enormous number of prior modes are retained; thus, a better basis is required.

7.4.2 LIS construction

Now we demonstrate the process of LIS construction using Algorithm 1, and the structure of the LIS under mesh refinement. To compute the LIS, we run Algorithm 1 for 500 iterations, using adaptive MALA [11] to simulate each subchain with length $L = 200$. We choose the truncation thresholds $\tau_{loc} = \tau_g = 0.1$. To explore the dimensionality and structure of the LIS versus mesh refinement, we carry out the same numerical experiment on a 60×20 coarse grid, a 120×40 intermediate grid, and a 180×60 fine grid. The dimension of the LIS versus number of iterations, the evolution of the convergence diagnostic (7.20), and the generalized eigenvalues after 500 iterations—for each level of grid refinement—are shown in Figure 7.4. Also, Figure 7.5 shows the first five LIS basis vectors for each level of discretization.

As shown in Figure 7.4(a), the dimension of the LIS changes rapidly in the first 100 iterations, then it stabilizes. Change in the dimension reflects the fact that the log-likelihood Hessian $H(x)$ varies locally in this non-Gaussian problem. We also observe that the 60×20 grid has a slightly larger final LIS dimension than the two refined grids: at the end of the adaptive construction, the LIS of the 60×20 grid has dimension 21, while the 120×40 grid and the 180×60 grid yield LIS dimensions of 20. This effect may be ascribed to larger discretization errors in the 60×20 grid.

The weighted distance (7.20) between each adjacent pair of likelihood-informed subspaces is used as the convergence diagnostic during the construction process. With any of the three discretizations, the weighted subspace

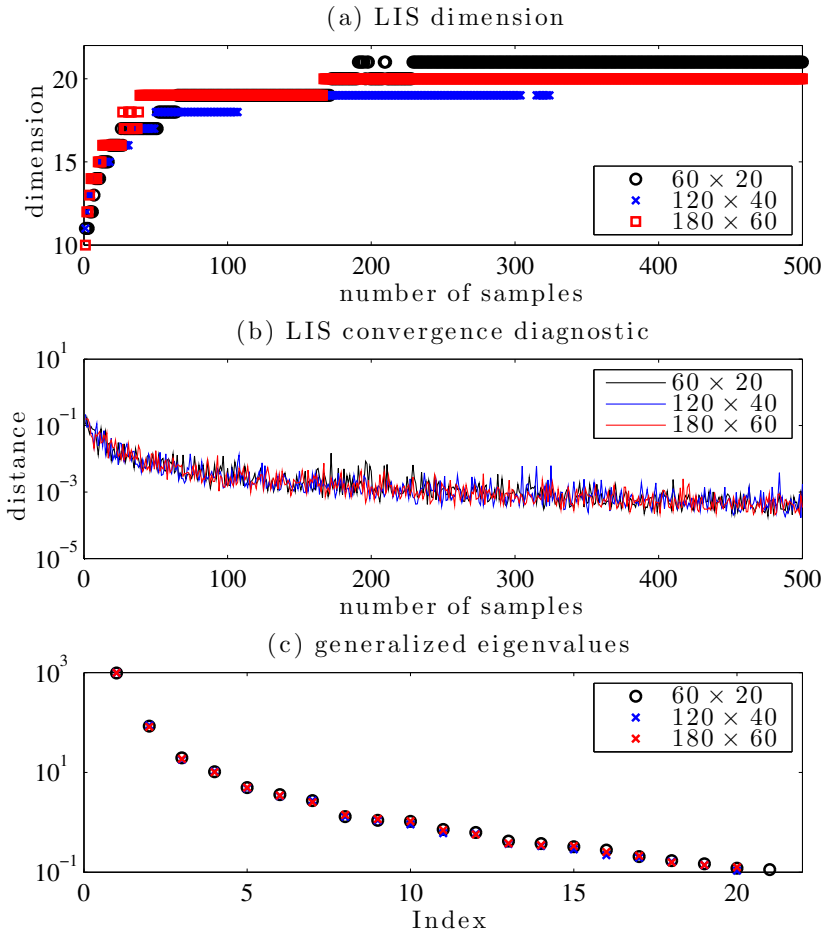

Figure 7.4: The dimension of the LIS and the convergence diagnostic (7.20) versus the number of samples used in the adaptive construction. Black, blue, and red markers represent the 60×20 grid, the 120×40 grid, and the 180×60 grid, respectively. Subplot (a) shows the dimension of the LIS; subplot (b) shows the weighted distance between successive subspaces; and subplot (c) shows the generalized eigenvalues $\gamma_i^{(k)}$ after $k = 500$ iterations.

distance at the end of adaptive construction is several orders of magnitude lower than at the beginning, as shown in Figure 7.4(b). We also observe that the rates of convergence of this diagnostic are comparable for all three levels of

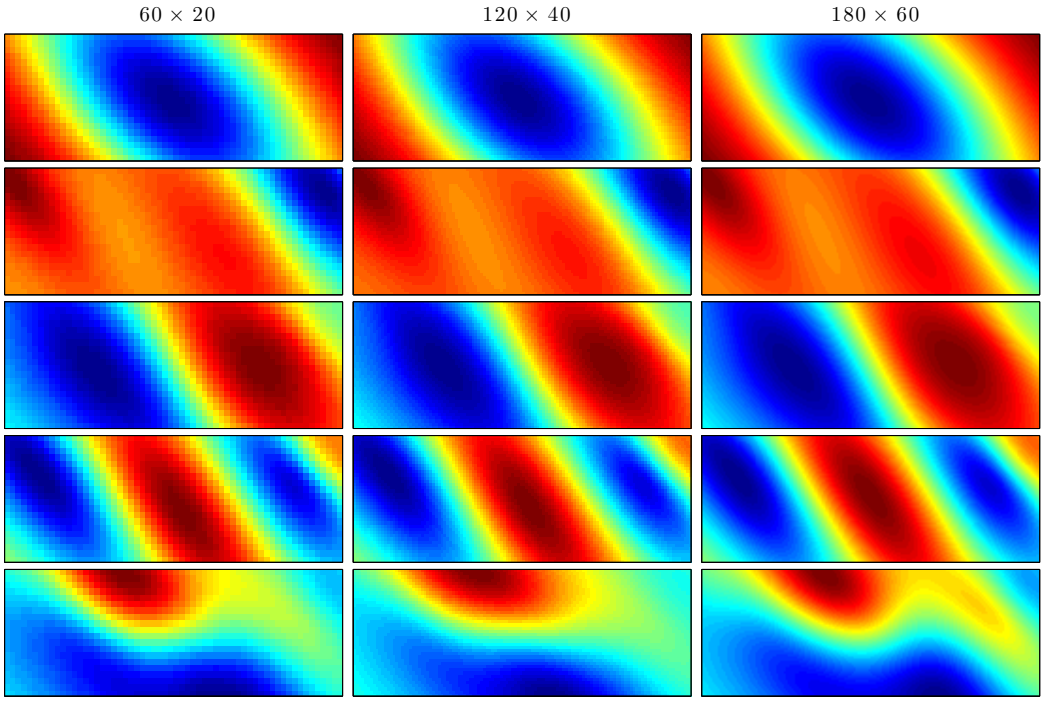

Figure 7.5: The first five LIS basis vectors (columns of Φ_5) for different levels of discretization of the inversion parameters x. In the figure, columns 1–3 correspond to the 60×20 grid, the 120×40 grid, and the 180×60 grid, respectively. The basis vectors in each column are ordered top to bottom by decreasing eigenvalue.

discretization. These figures suggest that while local variation of the Hessian is important in this problem (e.g., the dimension of the LIS doubles over the course of the iterations), much of this variation is well-explored after 100 or 200 iterations of Algorithm 1.

Since the forward model converges with grid refinement, we expect that the associated LIS should also converge. The generalized eigenvalues for all three grids are shown in Figure 7.4(c), where the spectra associated with all

three subspaces have very similar values. And as shown in Figure 7.5, the leading LIS basis vectors $\{\varphi_1, \ldots, \varphi_5\}$ have similar shapes for all three levels of grid refinement. Refinement leads to slightly more structure in φ_5, but the overall mode shapes are very close.

7.4.3 Estimation of the posterior mean and variance

With an LIS in hand, we apply the variance reduction procedure described in Section 7.3.3 to estimate the posterior mean and variance of the permeability field. Calculations in this subsection use the 120×40 discretization of the PDE and inversion parameters.

We first demonstrate the sampling performance of subspace MCMC, where we use adaptive MALA [11] to sample the LIS-defined reduced posterior $\tilde{\pi}(x_r|y)$ (7.14). We compare the results of subspace MCMC with the results of Hessian-preconditioned Langevin MCMC applied to the full posterior $\pi(x|y)$ (7.2) (referred to as "full-space MCMC" hereafter). The latter MCMC scheme results from an explicit discretization of the Langevin SDE, preconditioned by the inverse of the log-posterior Hessian evaluated at the posterior mode (see [56] for details). Note that we cannot precondition the full-dimensional Langevin SDE by the empirical posterior covariance as in adaptive MALA because of the high parameter dimension ($n = 4800$). In this setup, subspace MCMC and full-space MCMC require the same number of forward model and gradient evaluations for a given number of MCMC iterations.

To examine sampling performance, the autocorrelation of the log-likelihood

function and the autocorrelations of the parameters projected onto the first, third, and fifth LIS basis vectors are used as benchmarks. These results are shown in Figure 7.6. We run both algorithms for 10^6 iterations and discard the first half of the chains as burn-in. The top row of Figure 7.6 shows these benchmarks for both samplers. For all four benchmarks, subspace MCMC produces a faster decay of autocorrelation as a function of sample lag—i.e., a lower correlation between samples after any given number of MCMC steps.

Furthermore, as discussed in Section 7.3.3, even though the same number of forward model evaluations are required by subspace MCMC and full-space MCMC for a given number of samples, the computational cost of operations involving the square root of the prior covariance—used in sampling and evaluating the proposal distribution—can be much higher for full-space MCMC than subspace MCMC. In this test case, running subspace MCMC for 10^6 iterations cost 2.1×10^4 seconds of CPU time, while running full-space MCMC for the same number of iterations took 2.6×10^5 seconds. To incorporate this cost difference, the second row of Figure 7.6 shows the autocorrelation of the four benchmark quantities as a function of CPU time rather than sample lag. Here, we immediately observe that the autocorrelation per CPU time is further reduced by using subspace MCMC.

Of course, recall that to construct the LIS we simulated Algorithm 1 for 500 iterations. This costs roughly 2.2×10^4 seconds of CPU time, which is only 8.5% of the time required to run full-space MCMC for 10^6 steps. Therefore subspace MCMC, including the cost of LIS construction, takes less time to

produce a given number of samples than full-space MCMC *and* these samples are less correlated—i.e., of higher quality.

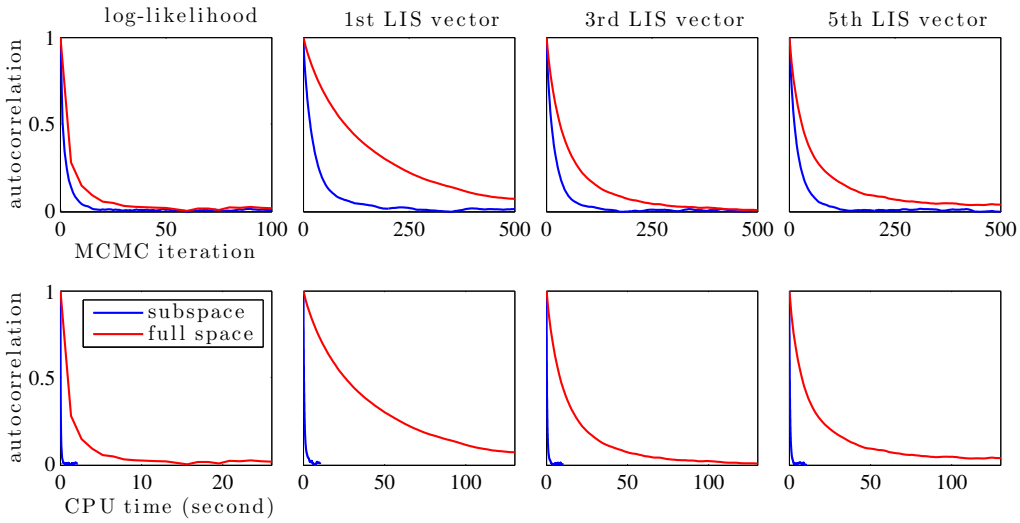

Figure 7.6: Autocorrelations of various benchmarks: blue line is subspace MCMC and red line is full-space MCMC. Column 1: log-likelihood function. Column 2: parameters projected onto the first LIS basis vector. Column 3: parameters projected onto the third LIS basis vector. Column 4: parameters projected onto the fifth LIS basis vector. Top row: Autocorrelation as a function of sample lag. Bottom row: Autocorrelation as a function of sample lag, where the latter is measured via CPU time.

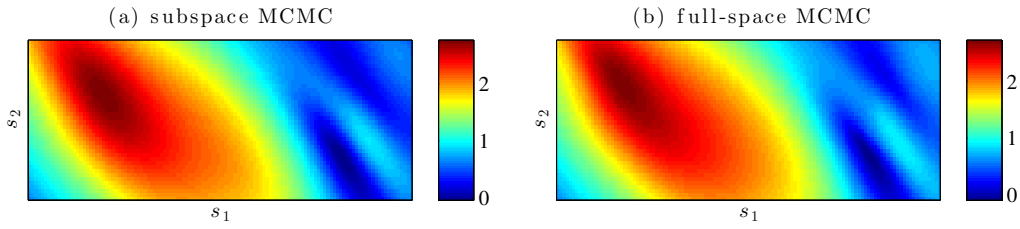

Figure 7.7: Estimates of posterior mean: (a) using subspace MCMC, (b) using full-space MCMC.

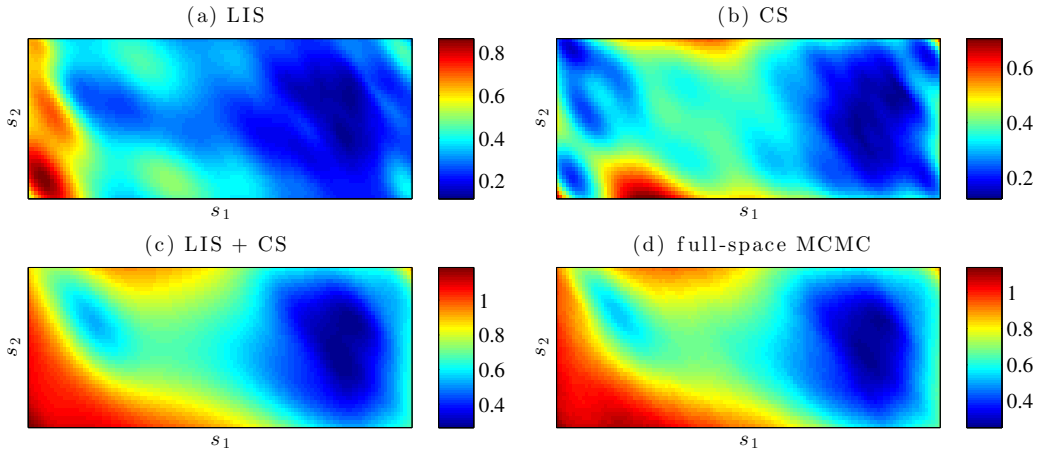

Figure 7.8: Estimation of the posterior variance: (a) empirical estimate using MCMC in the LIS; (b) analytical evaluation in the CS; (c) combined LIS + CS estimate; (d) for comparison, estimation using full-space MCMC.

We now compare reduced-variance estimates of the posterior mean and variance (obtained with subspace MCMC) with estimates obtained via full-space MCMC. The results are shown in Figures 7.7 and 7.8. Full-space MCMC and subspace MCMC yield very similar mean and variance estimates. Figures 7.8(a) and (b) distinguish the two components of the Rao-Blackwellized variance estimates described in Example 2. Variance in the LIS, shown in Figure 7.8(a), is estimated from MCMC samples, while variance in the CS, shown in Figure 7.8(b), is calculated analytically from the prior and the LIS projector. The sum of these two variance fields is shown in Figure 7.8(c), and it is nearly the same as the full-space result in Figure 7.8(d). In the central part of the domain where measurement sensors are not installed, we can observe that the variance is larger in the CS than in the LIS, and hence this part of the domain is prior-dominated. In the right part of the domain, the variance

is less prior-dominated, since this region is covered by observations.

7.4.4 The influence of data

The amount of information carried in the data affects the dimension and structure of the LIS. To demonstrate the impact of the data, we design a case study where different likelihood-informed subspaces are constructed under various observational scenarios. The same stationary groundwater problem defined in Section 7.4.1 is employed here. For the sake of computational efficiency, the problem domain $\Omega = [0, 3] \times [0, 1]$ is discretized by a slightly coarser 72×24 mesh. And to provide a stronger impulse to the groundwater system, the source/sink terms used in this example are different from those used in Sections 7.4.1–7.4.3. Along the boundary of the domain Ω, we evenly distribute a set of sources with a distance of 0.5 between the source centers. Two sinks are placed in the interior of the domain at locations $[0.5, 1]$ and $[2, 0.5]$. Each source has weight 1, while each sink has weight 3.5. We distributed sensors evenly over the domain $[0, 1] \times [0, 1] \cup [2, 3] \times [0, 1]$; starting with an inter-sensor spacing of 1/3, we incrementally refine the sensor distribution with spacings of 1/6, 1/12, and 1/24. This results in four different data sets, containing the noisy readings of 32, 98, 338, and 1250 sensors, respectively. The true permeability field, the sources/sinks, the simulated pressure field, and sensor distributions are shown in Figure 7.9.

As in Section 7.4.2, we run Algorithm 1 for 500 iterations to construct the LIS, using subchains of length $L = 200$. For data sets 1–4, the result-

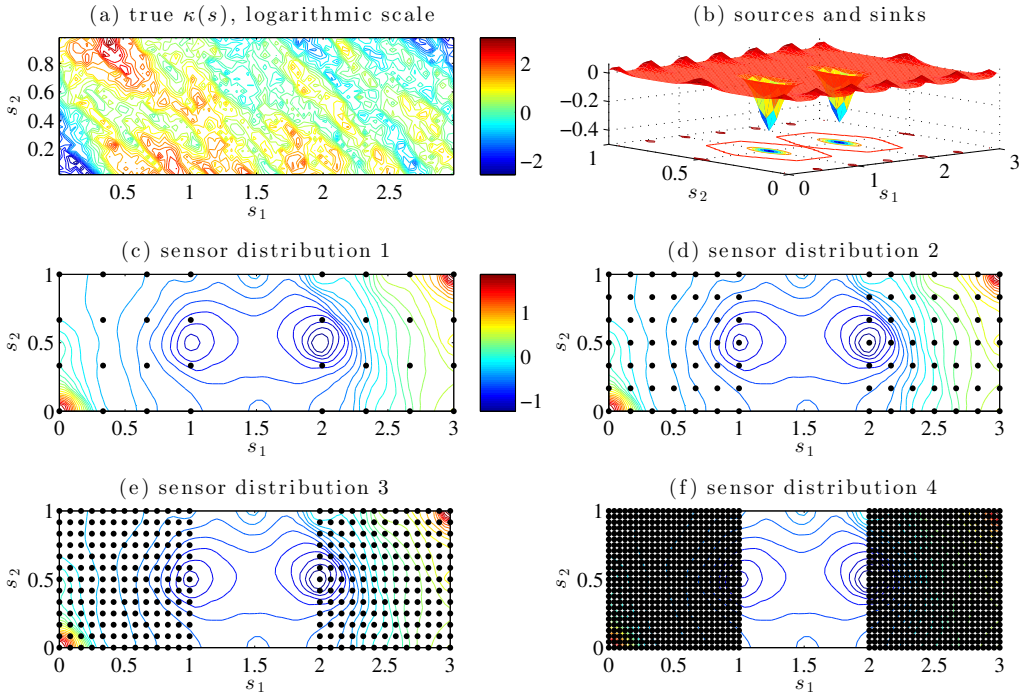

Figure 7.9: Setup of the elliptic inversion example for testing the influence of data. (a) True permeability field. (b) Sources and sinks. (c)–(f) Pressure field resulting from the permeability field defined in (a), and sensor distributions (black dots) for data sets 1–4.

ing LISs have dimension 24, 34, 50, and 83, respectively. The generalized eigenvalue spectrum for each LIS is shown in Figure 7.10. We note that the eigenvalues decay more slowly with increasing amounts of data. This behavior is expected; since the generalized eigenvalues reflect the impact of the likelihood, relative to the prior, more data should lead to more directions where the likelihood dominates the prior.

Since the sensors for all four data sets occupy the same area of the

spatial domain, we expect that the four likelihood-informed subspaces should share a similar low frequency structure. However, the high frequency structures in each LIS might differ from each other under refinement of the sensor distribution. Thus the LIS basis vectors corresponding to the largest eigenvalues should share a similar pattern, while the LIS basis vectors corresponding to the relatively small eigenvalues might have different patterns. We observe this effect in the numerical experiments carried out here; Figure 7.11 shows the first and fifteenth LIS basis vector for each of the data sets.

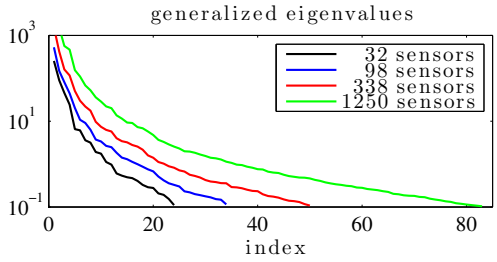

Figure 7.10: Generalized eigenvalues associated with the likelihood-informed subspace under refinement of the observations. The black, blue, red, and green lines show eigenvalues for data sets 1–4, with 32, 98, 338, and 1250 sensors, respectively.

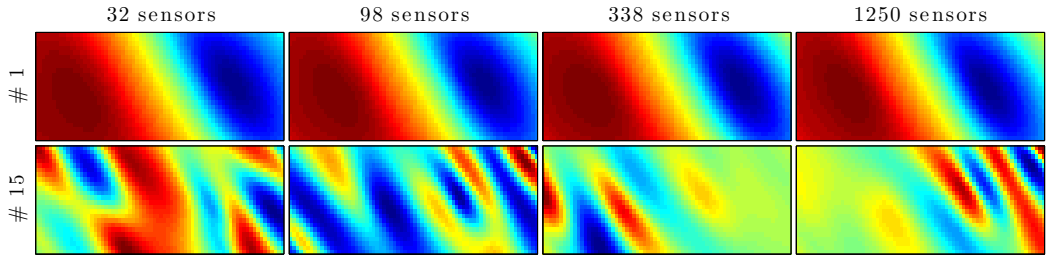

Figure 7.11: The first and fifteenth LIS basis vectors for each of the four data sets.

7.5 Example 2: atmospheric remote sensing

In this section, we apply the dimension reduction approach to a realistic atmospheric satellite remote sensing problem. The problem is to invert the concentrations of various gases in the atmosphere using the measurement system applied in the GOMOS satellite instrument, which stands for *Global Ozone MOnitoring System*.

GOMOS is an instrument on board ESA's Envisat satellite, and was operational for about 10 years before the connection with the satellite was lost in May 2012. The GOMOS instrument performs so-called star occultation measurements; it measures, at different wavelengths, the absorption of starlight as it travels through the atmosphere. Different gases in the atmosphere (such as ozone, nitrogen dioxide and aerosols) leave fingerprints in the measured intensity spectra. The task of the inversion algorithm is to infer the concentrations of these gases based on the measurements.

The GOMOS inverse problem is known to be ill-posed; the intensity spectra may contain strong information about the major gases (like O_3) at some altitudes, whereas some minor gases (like aerosols) at some altitudes may be practically unidentifiable and totally described by the prior. Thus, the GOMOS problem is a good candidate for our approach: the dimension of the likelihood informed subspace is expected to be small and the prior contribution large.

Next, we briefly present the GOMOS theory and the inverse problem

setup. For more details about the GOMOS instrument and the Bayesian treatment of the inverse problem, see [94] and the references therein.

7.5.1 The GOMOS model

The GOMOS instrument repeatedly measures light intensities I_λ at different wavelengths λ. First, a reference intensity spectrum I_{ref} is measured above the atmosphere. The so-called transmission spectrum is defined as $T_\lambda = I_\lambda / I_{\text{ref}}$. The transmissions measured at wavelength λ along the ray path l are modelled using Beer's law:

$$T_{\lambda,l} = \exp\left(- \int_l \sum_{\text{gas}} \alpha_\lambda^{\text{gas}}(z(s)) \rho^{\text{gas}}(z(s)) ds \right), \tag{7.25}$$

where $\rho^{\text{gas}}(z(s))$ is the density of a gas (unknown parameter) at tangential height z. The so called cross-sections $\alpha_\lambda^{\text{gas}}$, known from laboratory measurements, define how much a gas absorbs light at a given wavelength.

To approximate the integrals in (7.25), the atmosphere is discretized. The geometry used for inversion resembles an onion: the gas densities are assumed to be constant within spherical layers around the Earth. The GOMOS measurement principle is illustrated in Figure 7.12 below.

Here, we assume that the cross-sections do not depend on height. In the inverse problem we have n_{gas} gases, n_λ wavelengths, and the atmosphere is divided into n_{alts} layers. The discretisation is fixed so that number of measurement lines is equal to the number of layers. Approximating the integrals by sums in the chosen grid, and combining information from all lines and all

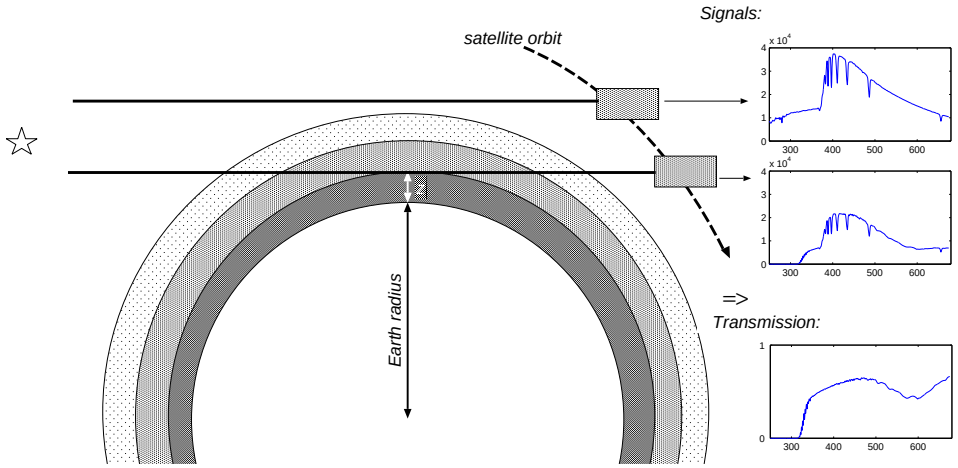

Figure 7.12: The principle of the GOMOS measurement. The reference intensity is measured above the atmosphere. The observed transmission spectrum is the attenuated spectrum (measured through the atmosphere) divided by the reference spectrum. The atmosphere is presented locally as spherical layers around the Earth. Note that the thickness of the layers is much larger relative to the Earth in this figure than in reality. The figure is adopted from [94], with the permission of the authors.

wavelengths, we can write the model in matrix form as follows:

$$T = \exp(-CB^\top A^\top), \tag{7.26}$$

where $T \in \mathbb{R}^{n_\lambda \times n_{\text{alts}}}$ are the modelled transmissions, $C \in \mathbb{R}^{n_\lambda \times n_{\text{gas}}}$ contains the cross-sections, $B \in \mathbb{R}^{n_{\text{alts}} \times n_{\text{gas}}}$ contains the unknown densities and $A \in \mathbb{R}^{n_{\text{alts}} \times n_{\text{alts}}}$ is the geometry matrix that contains the lengths of the lines of sight in each layer.

Computationally, it is convenient to deal with vectors of unknowns. We vectorize the above model using the identity $\text{vec}(CB^\top A^\top) = (A \otimes C)\text{vec}(B^\top)$, where $\otimes$ denotes the Kronecker product and vec is the standard vectorization

obtained by stacking the columns of the matrix argument on top of each other. Thus, the likelihood model is written in vector form as follows:

$$y = \text{vec}(T) + e = \exp\left(-(A \otimes C)\text{vec}(B^\top)\right) + e, \tag{7.27}$$

where e is the measurement error, for which we apply an independent Gaussian model with known variances.

Note that, in principle, the model (7.27) could be linearized by taking logarithms of both sides, which is usually done for such tomography problems (like X-ray computerized tomography). For this problem, linearisation can cause problems, since the signal from the star is often smaller compared to the background noise in the measurement.

7.5.2 Data and prior

Here, we generate synthetic data by solving the forward model (7.27) with known gas densities x. In the example, we have 4 gas profiles to be inverted. The atmosphere is discretized into 50 layers, and the total dimension of the problem is thus 200. The simulated data are illustrated in Figure 7.13.

We estimate the log-profiles $x = \log(\text{vec}(B^\top))$ of the gases instead of the densities B directly. We set a Gaussian process prior for the profiles, which yields $x_i \sim N(\mu_i, \Sigma_i)$, where x_i denotes the elements of vector x corresponding to gas i. The elements of the 50×50 covariance matrices are calculated based on the squared exponential covariance function

$$C_i(s, s') = \sigma_i \exp(-(s - s')^2/2s_{0,i}^2), \tag{7.28}$$

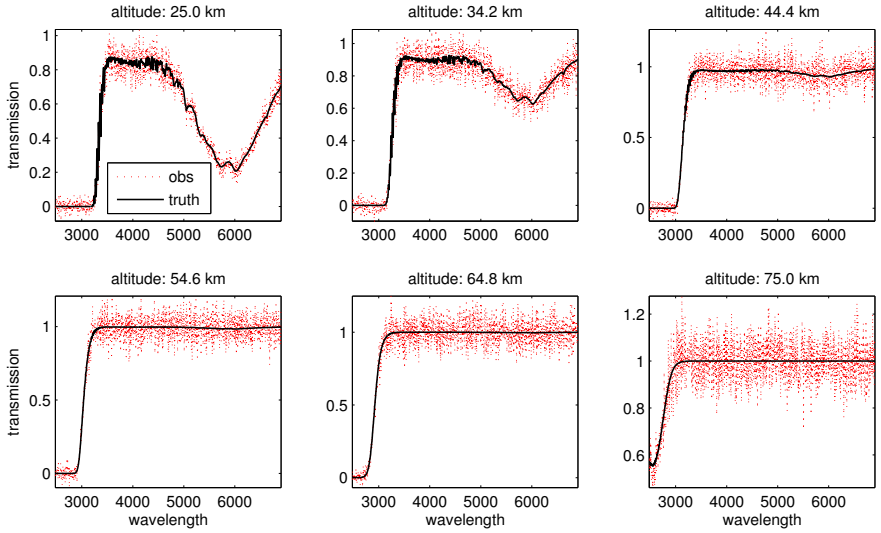

Figure 7.13: GOMOS example setup: the true transmissions (black) and the observed transmissions (red) for 6 altitudes.

where the parameter values are $\sigma_1 = 5.22$, $\sigma_2 = 9.79$, $\sigma_3 = 23.66$, $\sigma_4 = 83.18$, and $s_{0,i} = 10$ for all i. The priors are chosen to promote smooth profiles and to give a rough idea about the magnitude of the density values. The prior is illustrated in Figure 7.14.

7.5.3 Inversion results

In this particular synthetic example, we know that gas 1 is very well identified by the data. The data also contain information about gases 2 and 3 at some altitudes. Gas 4, on the other hand, is totally unidentified by the data.

The LIS is constructed using 200 samples—i.e., 200 iterations of Al-

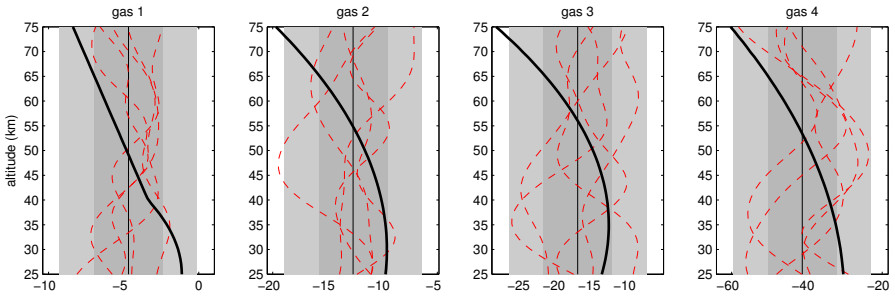

Figure 7.14: True log-profiles for the 4 gases (black solid lines), 50% and 95% confidence envelopes for the prior (grey areas) and 5 samples from the prior (red dashed lines).

gorithm 1—starting with the Hessian at the posterior mode. The subspace convergence diagnostic and the generalized eigenvalues are shown in Figure 7.15. We choose the truncation thresholds $\tau_{loc} = \tau_g = 0.5$. The dimension of the LIS in the end was 22.

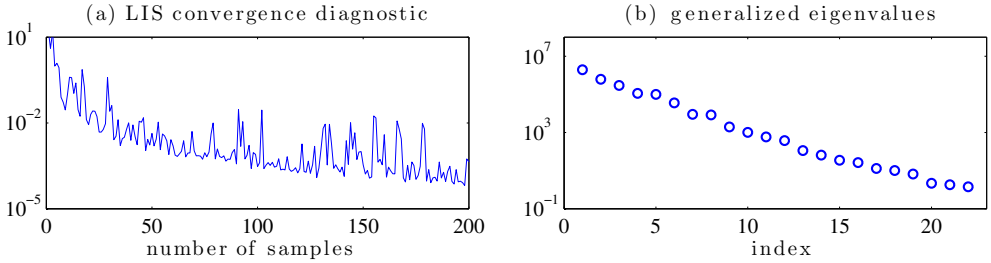

Figure 7.15: Left: the convergence diagnostic (7.20) versus the number of samples used to construct the LIS. Right: the generalized eigenvalues associated with the final LIS.

We compute 10^6 samples in both the LIS and in the full 200-dimensional space using the Hessian-preconditioned MALA algorithm. In Figure 7.16, the first two columns show the mean gas profile and the mean $\pm$ 1 and 2 standard deviations in the LIS and in the complement space (CS). The third column

shows the combined posterior from the LIS and the CS; for comparison, results from full-space MCMC are shown in the fourth column. Note the different scales on the horizontal axes throughout the figure. We observe that the subspace approach, where MCMC is applied only in a 22-dimensional space, yields results very similar to those of full MCMC. In addition, comparing the contributions of the LIS and CS indicates that gas 1 is dominated by the likelihood, whereas the posterior distribution of gas 4 is entirely determined by the prior. Note that the CS contribution for gas 1 is tiny (check the scale), while the LIS contribution for gas 4 is also very small. For gases 2 and 3, the lower altitudes are likelihood-dominated, while the higher altitudes have more contribution from the prior. The full-space MCMC results for gas 4 show a slightly non-uniform mean, but this appears to be the result of sampling variance. By avoiding sampling altogether in the CS, the subspace approach most likely yields a more accurate posterior in this case.

To further illustrate the approach, we plot the first six basis vectors of the LIS in Figure 7.17. One can see that the first basis vectors mainly include features of gas 1, which is most informed by the data. The first basis vectors also contain some features of gases 2 and 3 in lower altitudes. Gas 4 is not included in the LIS at all.

The dimension reduction obtained via the subspace approach is expected to yield better mixing than the full-space MCMC. For the GOMOS case, the chain autocorrelations for subspace and full-space MCMC are compared in Figure 7.18. The subspace sampler shows much faster decay of the

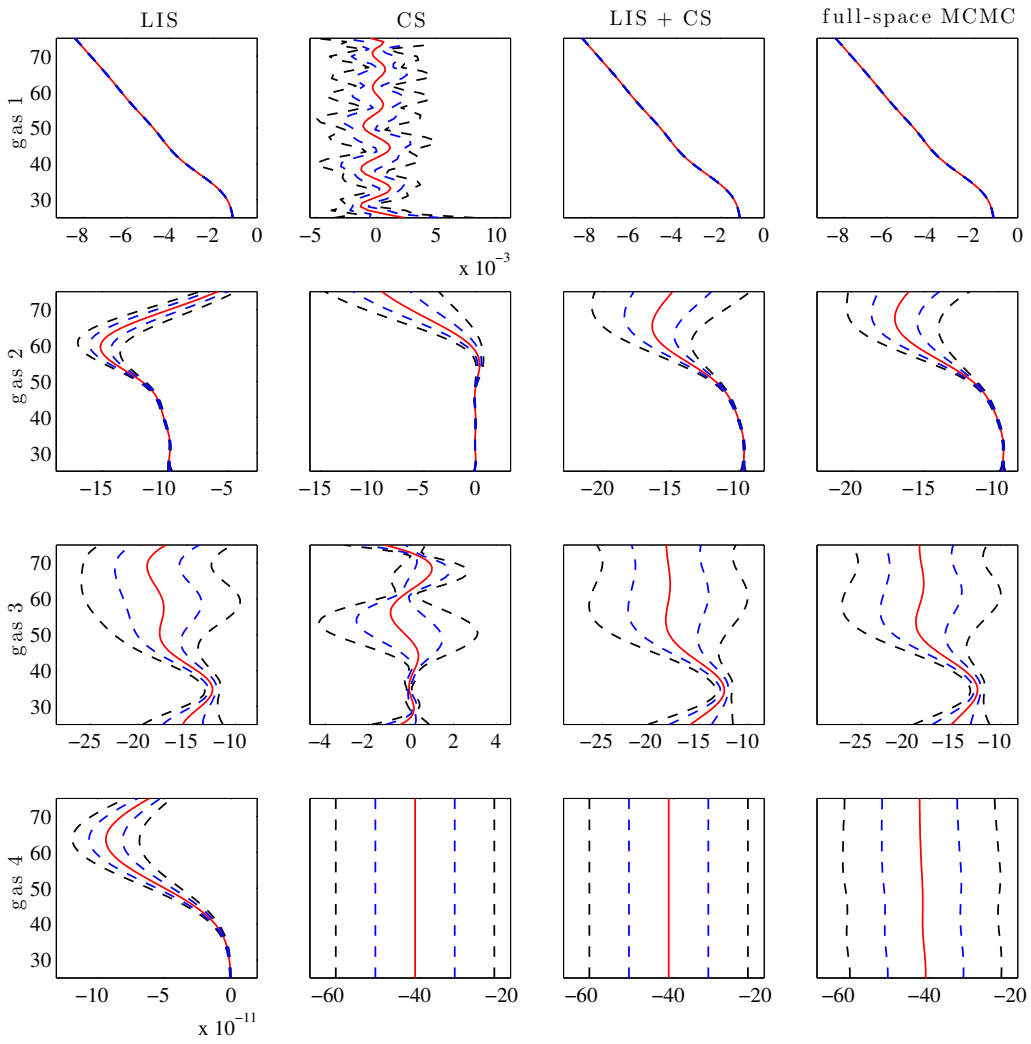

Figure 7.16: Mean and $\pm 1/\pm 2$ standard deviations for the 4 gas profiles computed from the LIS samples alone (1st column), CS alone (2nd column) and when the LIS and CS are combined (3rd column). The same quantities computed from full-space MCMC are given in the 4th column.

autocorrelations than full-space MCMC.

In this test case, the subspace MCMC also has lower computational

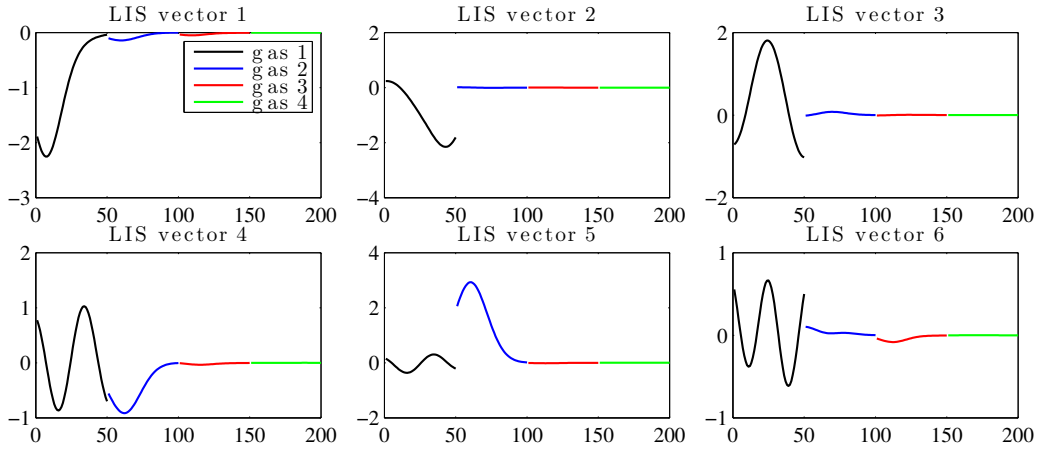

Figure 7.17: The first six LIS basis vectors for the remote sensing example. The colors indicate the components of the unknown vector corresponding to the different gases. In each subfigure, the x-axis denotes the index of the parameter vector, and, for each gas, the components are ordered from low altitudes to high altitudes. (For example, the black line in each figure shows gas 1 profiles from low altitudes to high altitudes, etc.)

cost compared to full-space MCMC. To simulate a Markov chain for 10^6 iterations, the subspace MCMC consumed about 2560 seconds of CPU time, while the full-space MCMC cost 3160 CPU seconds. We note that the CPU time reduction is not as significant as the elliptic example, because the prior covariance is a 200×200 dimensional matrix, which is much smaller than the covariance matrix used in the elliptic example. To construct the LIS, we simulated Algorithm 1 for 200 iterations. This cost about 136 seconds of CPU time, which is only about 4.3% of the CPU time used to run full-space MCMC for 10^6 steps.

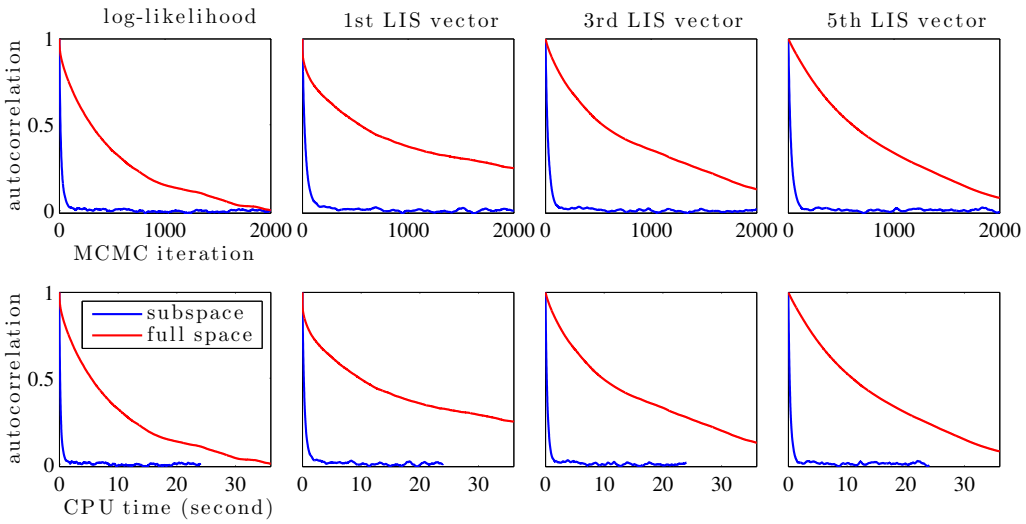

Figure 7.18: Autocorrelations of full-space (red) and subspace (blue) MCMC for the log-likelihood (1st column) and for the samples projected onto the first, third, and fifth LIS basis vectors (2nd, 3rd and 4rd columns). Top rows shows the autocorrelations computed per MCMC step and bottom row per CPU time.

7.6 Conclusions

In this paper, we present a new approach for dimension reduction in nonlinear inverse problems with Gaussian priors. Our approach is based on dividing the parameter space into two subspaces: a likelihood-informed subspace (LIS) where the likelihood has a much greater influence on the posterior than the prior distribution, and the complement to the LIS where the Gaussian prior dominates. We explore the posterior projected onto the LIS (the "difficult" and non-Gaussian part of the problem) with Markov chain Monte Carlo while treating the complement space as exactly Gaussian. This approximation allows us to analytically integrate many functions over the complement space when

estimating their posterior expectations; the result is a Rao-Blackwellization or de-randomization procedure that can greatly reduce the variance of posterior estimates. Particularly in inverse problems—where information in the data is often limited and the solution of the problem relies heavily on priors—the dimension of the LIS is expected to be small, and the majority of the directions in the parameter space can be handled analytically.

The dimension reduction approach is based on theory developed for the linear case; in [186] it is shown that in linear-Gaussian problems, the eigendecomposition of the prior-preconditioned log-likelihood Hessian yields an optimal low-rank update from the prior to the posterior, which can be interpreted in terms of a projector whose range is the LIS. Here, we generalize the approach to nonlinear problems, where the log-likelihood Hessian varies over the parameter space. Our solution is to construct many local likelihood-informed subspaces over the support of the posterior and to combine them into a single global LIS. We show how the global LIS can be constructed efficiently in an adaptive manner, starting with the LIS computed at the posterior mode and iteratively enriching the global LIS until a weighted subspace convergence criterion is met.

We demonstrate the approach with two numerical examples. First is an elliptic PDE inverse problem, based on a simple model of subsurface flow. Though the dimension of the parameter space in our experiments ranges from 1200 to 10800, the dimension of the LIS remains only around 20 and is empirically discretization-invariant. Exploring the LIS by MCMC and analytically

treating the Gaussian complement produces mean and variance fields very similar to those computed via MCMC in the full space. Yet the mixing properties and the computational cost of MCMC in the LIS are dramatically improved over those of full-space MCMC. Our second demonstration is an atmospheric remote sensing problem, where the goal is to infer the concentrations of chemical species in the atmosphere using star occultation measurements, as on the satellite-borne GOMOS instrument. The dimension of the full problem used here was 200 (four gaseous species and 50 altitudes for each), while the dimension of the LIS was 22. Again, dimension reduction significantly improves the mixing properties of MCMC without sacrificing accuracy.

To conclude, our dimension reduction approach appears to offer an efficient way to probe and exploit the structure of nonlinear inverse problems in order to perform Bayesian inference at a large scale, where standard algorithms are plagued by the curse of dimensionality. The approach also opens up interesting further research questions: it may be useful, for instance, to apply reduced-order and surrogate modeling techniques in the LIS, making them applicable to much larger problems than before.

7.7 Acknowledgements

We acknowledge Marko Laine and Johanna Tamminen from the Finnish Meteorological Institute for providing us with the GOMOS figure and codes that served as the baseline for our implementation for the remote sensing example. This work was supported by the US Department of Energy, Office of

Advanced Scientific Computing (ASCR), under grant numbers DE-SC0003908 and DE-SC0009297.

Bibliography

[1] Performance applications programming interface (PAPI).

[2] P. Abrahamsen. A review of Gaussian random fields and correlation functions. Technical Report 2nd edition, Norwegian Computing Center, Box 114, Blindern N-0314 Oslo, Norway, 1997.

[3] V. Akçelik, J. Bielak, G. Biros, I. Epanomeritakis, A. Fernandez, O. Ghattas, E. J. Kim, J. Lopez, D. R. O'Hallaron, T. Tu, and J. Urbanic. High resolution forward and inverse earthquake modeling on terascale computers. In *SC03: Proceedings of the International Conference for High Performance Computing, Networking, Storage, and Analysis*. ACM/IEEE, 2003. Gordon Bell Prize for Special Achievement.

[4] V. Akçelik, G. Biros, and O. Ghattas. Parallel multiscale Gauss-Newton-Krylov methods for inverse wave propagation. In *Proceedings of IEEE/ACM SC2002 Conference*, Baltimore, MD, Nov. 2002. SC2002 Best Technical Paper Award.

[5] V. Akçelik, G. Biros, O. Ghattas, J. Hill, D. Keyes, and B. van Bloeman Waanders. Parallel PDE-constrained optimization. In M. Heroux, P. Raghaven, and H. Simon, editors, *Parallel Processing for Scientific Computing*. SIAM, 2006.

[6] K. Aki and P. G. Richards. *Quantitative Seismology*. University Science Books, 2nd edition, 2002.

[7] C. Andrieu and E. Moulines. On the ergodicity properties of some adaptive MCMC algorithms. *The Annals of Applied Probability*, 16(3):1462–1505, 2006.

[8] A. Apte, M. Hairer, A. M. Stuart, and J. Voss. Sampling the posterior: An approach to non-gaussian data assimilation. *Physica D: Nonlinear Phenomena*, 2006.

[9] T. Arbogast and J. L. Bona. *Methods of Applied Mathematics*. University of Texas at Austin, 2008. Lecture notes in applied mathematics.

[10] S. R. Arridge, J. P. Kaipio, V. Kolehmainen, M. Schweiger, E. Somersalo, T. Tarvainen, and M. Vauhkonen. Approximation errors and model reduction with an application in optical diffusion tomography. *Inverse Problems*, 22(1):175, 2006.

[11] Y. F. Atchade. An adaptive version for the Metropolis adjusted Langevin algorithm with a truncated drift. *Methodology and Computing in Applied Probability*, 8(2):235–254, 2006.

[12] H. Auvinen, J. M. Bardsley, H. Haario, and T. Kauranne. Large-scale Kalman filtering using the limited memory BFGS method. *Electronic Transactions on Numerical Analysis*, 35:217–233, 2009.

[13] H. Auvinen, J. M. Bardsley, H. Haario, and T. Kauranne. The variational Kalman filter and an efficient implementation using limited memory BFGS. *International Journal for Numerical Methods in Fluids*, 64(3):314–335, 2010.

[14] V. A. Badri Narayanan and N. Zabaras. Stochastic inverse heat conduction using a spectral approach. *International Journal for Numerical Methods Engineering*, 60(9):1569–1593, 2004.

[15] Z. Bai, J. Demmel, J. Dongarra, A. Ruhe, and H. van der Vorst. *Templates for the Solution of Algebraic Eigenvalue Problems: A Practical Guide*, volume 11. SIAM, 2000.

[16] J. M. Bardsley. Gaussian Markov random field priors for inverse problems. *Inverse Problems and Imaging*, 2013.

[17] R. Barrett, M. W. Berry, T. F. Chan, J. Demmel, J. Donato, J. Dongarra, V. Eijkhout, R. Pozo, C. Romine, and H. Van der Vorst. *Templates for the Solution of Linear Systems: Building Blocks for Iterative Methods*, volume 43. SIAM, 1994.

[18] E. B. Becker, G. F. Carey, and J. T. Oden. *Finite Elements: An Introduction, Vol I.* Prentice Hall, Englewoods Cliffs, New Jersey, 1981.

[19] A. Beskos, G. Roberts, and A. Stuart. Optimal scalings for local Metropolis-Hastings chains on nonproduct targets in high dimensions. *The Annals of Applied Probability*, 19(3):863–898, 2009.

[20] A. Beskos, G. O. Roberts, A. M. Stuart, and J. Voss. MCMC methods for diffusion bridges. *Stochastic Dynamics*, 8(3):319–350, 2008.

[21] G. Biros and O. Ghattas. Parallel Lagrange–Newton–Krylov–Schur methods for PDE–constrained optimization. Part II: The Lagrange–Newton solver and its application to optimal control of steady viscous flows. *SIAM Journal on Scientific Computing*, 27(2):714–739, 2005.

[22] A. Bonito, R. H. Nochetto, and S. M. Pauletti. Geometrically consistent mesh modification. *SIAM Journal on Numerical Analysis*, 48(5):1877–1899, 2010.

[23] A. Borzì and V. Schulz. *Computational Optimization of Systems Governed by Partial Differential Equations*. SIAM, 2012.

[24] S. Brooks and A. Gelman. General methods for monitoring convergence of iterative simulations. *Journal of Computational and Graphical Statistics*, 7:434–456, 1998.

[25] S. Brooks, A. Gelman, G. Jones, and X. L. Meng, editors. *Handbook of Markov Chain Monte Carlo*. Taylor & Francis, 2011.

[26] T. Bui-Thanh, C. Burstedde, O. Ghattas, J. Martin, G. Stadler, and L. C. Wilcox. Extreme-scale UQ for Bayesian inverse problems governed by PDEs. In *SC12: Proceedings of the International Conference for High Performance Computing, Networking, Storage and Analysis*, 2012.

[27] © 2012 IEEE. Reprinted, with permission, from T. Bui-Thanh, C. Burstedde, O. Ghattas, J. Martin, G. Stadler, and L. C. Wilcox. Extreme-scale UQ for Bayesian inverse problems governed by PDEs. In *SC12: Proceedings of the International Conference for High Performance Computing, Networking, Storage and Analysis*, November 2012.

[28] T. Bui-Thanh and O. Ghattas. Analysis of an *hp*-non-conforming discontinuous Galerkin spectral element method for wave propagation. *SIAM Journal on Numerical Analysis*, 50(3):1801–1826, 2012.

[29] T. Bui-Thanh and O. Ghattas. Analysis of the Hessian for inverse scattering problems. Part I: Inverse shape scattering of acoustic waves. *Inverse Problems*, 28(5):055001, 2012.

[30] T. Bui-Thanh and O. Ghattas. Analysis of the Hessian for inverse scattering problems. Part II: Inverse medium scattering of acoustic waves. *Inverse Problems*, 28(5):055002, 2012.

[31] T. Bui-Thanh and O. Ghattas. Analysis of the Hessian for inverse scattering problems. Part III: Inverse medium scattering of electromagnetic waves. Inverse Problems and Imaging, 2013.

[32] T. Bui-Thanh, O. Ghattas, and D. Higdon. Adaptive Hessian-based nonstationary Gaussian process response surface method for probability density approximation with application to Bayesian solution of large-scale inverse problems. *SIAM Journal on Scientific Computing*, 34(6):A2837–A2871, 2012.

[33] T. Bui-Thanh, O. Ghattas, J. Martin, and G. Stadler. A computational framework for infinite-dimensional Bayesian inverse problems Part I: The linearized case, with application to global seismic inversion. *SIAM Journal on Scientific Computing*, 35(6):A2494–A2523, 2013.

[34] T. Bui-Thanh, K. Willcox, and O. Ghattas. Parametric reduced-order models for probabilistic analysis of unsteady aerodynamic applications. *AIAA Journal*, 46:2520–2529, 2008.

[35] C. Burstedde and O. Ghattas. Algorithmic strategies for full waveform inversion: 1D experiments. *Geophysics*, 74(6):WCC37–WCC46, 2009.

[36] C. Burstedde, O. Ghattas, M. Gurnis, T. Isaac, G. Stadler, T. Warburton, and L. C. Wilcox. Extreme-scale AMR. In *SC10: Proceedings of the International Conference for High Performance Computing, Networking, Storage and Analysis*. ACM/IEEE, 2010.

[37] C. Burstedde, L. C. Wilcox, and O. Ghattas. `p4est`: Scalable algorithms for parallel adaptive mesh refinement on forests of octrees. *SIAM Journal on Scientific Computing*, 33(3):1103–1133, 2011.

[38] D. Calvetti. Preconditioned iterative methods for linear discrete ill-posed problems from a Bayesian inversion perspective. *Journal of computational and applied mathematics*, 198(2):378–395, 2007.

[39] D. Calvetti, B. Lewis, and L. Reichel. On the regularizing properties of the gmres method. *Numerische Mathematik*, 91(4):605–625, 2002.

[40] D. Calvetti, D. McGivney, and E. Somersalo. Left and right preconditioning for electrical impedance tomography with structural information. *Inverse Problems*, 28(5):055015, 2012.

[41] D. Calvetti and E. Somersalo. Priorconditioners for linear systems. *Inverse problems*, 21(4):1397, 2005.

[42] B. Carlin and T. Louis. *Bayesian Methods for Data Analysis*. Chapman & Hall/CRC, 3rd edition, 2009.

[43] L. Carrington, D. Komatitsch, M. Laurenzano, M. M. Tikir, D. Michéa, N. L. Goff, A. Snavely, and J. Tromp. High-frequency simulations of global seismic wave propagation using SPECFEM3D GLOBE on 62K processors. In *SC08: Proceedings of the International Conference for High Performance Computing, Networking, Storage, and Analysis*. ACM/IEEE, 2008.

[44] G. Casella and C. P. Robert. Rao-Blackwellisation of sampling schemes. *Biometrika*, 83(1):81–94, 1996.

[45] J. Chen, M. Anitescu, and Y. Saad. Computing $f(A)b$ via least squares polynomial approximations. *SIAM Journal on Scientific Computing*, 33(1):195–222, Feb. 2011.

[46] J. A. Christen and C. Fox. Markov chain Monte Carlo using an approximation. *Journal of Computational & Graphical Statistics*, 14(4):795–810, December 2005.

[47] J. A. Christen and C. Fox. A general purpose sampling algorithm for continuous distributions (the t-walk). *Bayesian Analysis*, 5(2):263–283, 2010.

[48] R. Christensen. *Plane Answers to Complex Questions: The Theory of Linear Models*. Springer-Verlag, 1987.

[49] J. Chung and M. Chung. An efficient approach for computing optimal low-rank regularized inverse matrices. *Inverse Problems*, 30(11):114009, 2014.

[50] J. Chung, M. Chung, and D. P. O'Leary. Optimal regularized low rank inverse approximation. *Linear Algebra and its Applications*, 468:260 – 269, 2015.

[51] D. Colton and R. Kress. *Inverse Acoustic and Electromagnetic Scattering*. Applied Mathematical Sciences, Vol. 93. Springer-Verlag, Berlin, Heidelberg, New-York, Tokyo, second edition, 1998.

[52] S. L. Cotter, M. Dashti, J. C. Robinson, and A. M. Stuart. Bayesian inverse problems for functions and applications to fluid mechanics. *Inverse Problems*, 25(11):115008, 2009.

[53] S. L. Cotter, M. Dashti, and A. M. Stuart. Approximation of Bayesian inverse problems for PDEs. *SIAM Journal on Numerical Analysis*, 48(1):322–345, 2010.

[54] S. L. Cotter, G. O. Roberts, A. M. Stuart, and D. White. MCMC methods for functions: modifying old algorithms to make them faster. *Statistical Science*, 28:424–446, 2013.

[55] T. Cui, C. Fox, and M. J. O'Sullivan. Adaptive error modelling in MCMC sampling for inverse problems. *Journal of the Royal Statistical Society: Series B*, 2012. Submitted.

[56] T. Cui, K. J. H. Law, and Y. M. Marzouk. Dimension–independent likelihood–informed MCMC. *Preprint*, 2014. arXiv preprint arXiv:1411.3688.

[57] T. Cui, J. Martin, Y. M. Marzouk, A. Solonen, and A. Spantini. Likelihood-informed dimension reduction for nonlinear inverse problems. *Inverse Problems*, 30(11):114015, 2014.

[58] Y. Cui, K. B. Olsen, T. H. Jordan, K. Lee, J. Zhou, P. Small, D. Roten, G. Ely, D. K. Panda, A. Chourasia, J. Levesque, S. M. Day, and P. Maechling. Scalable earthquake simulation on petascale supercomputers. In *SC10: Proceedings of the International Conference for High Performance Computing, Networking, Storage, and Analysis*. ACM/IEEE, 2010.

[59] A. Demlow. Higher-order finite element methods and pointwise error estimates for elliptic problems on surfaces. *SIAM Journal on Numerical Analysis*, 47(2):805–827, 2009.

[60] C. R. Dietrich and G. N. Newsam. Fast and exact simulation of stationary gaussian processes through circulant embedding of the covariance

matrix. *SIAM Journal on Scientific Computing*, 18(4):1088–1107, 1997.

[61] P. Dostert, Y. Efendiev, T. Y. Hou, and W. Luo. Coarse-gradient Langevin algorithms for dynamic data integration and uncertainty quantification. *Journal of Computational Physics*, 217:123–142, 2006.

[62] L. Dykes and L. Reichel. Simplified gsvd computations for the solution of linear discrete ill-posed problems. *Journal of Computational and Applied Mathematics*, 255:15–27, 2014.

[63] A. M. Dziewonski and D. L. Anderson. Preliminary reference earth model. *Physics of the Earth and Planetary Interiors*, 25(4):297–356, 1981.

[64] G. Dziuk. Finite elements for the Beltrami operator on arbitrary surfaces. In S. Hildebrandt and R. Leis, editors, *Partial Differential Equations and Calculus of Variations*, volume 1357 of *Lecture Notes in Mathematics*, pages 142–155. Springer Berlin Heidelberg, 1988.

[65] C. Eckart and G. Young. The approximation of one matrix by another of lower rank. *Psychometrika*, 1(3):211–218, 1936.

[66] Y. Efendiev, T. Hou, and W. Luo. Preconditioning Markov Chain Monte Carlo simulations using coarse-scale models. *SIAM Journal on Scientific Computing*, 28(2):776–803, 2006.

[67] I. Epanomeritakis, V. Akçelik, O. Ghattas, and J. Bielak. A Newton-CG method for large-scale three-dimensional elastic full-waveform seismic inversion. *Inverse Problems*, 24(3):034015 (26pp), 2008.

[68] L. C. Evans. *Partial Differential Equations*. American Mathematical Society, Providence, RI, 1998.

[69] G. Evensen. *Data Assimilation*. Springer, 2007.

[70] A. Fichtner. *Full seismic waveform modelling and inversion*. Springer, 2011.

[71] A. Fichtner, H. Igel, H.-P. Bunge, and B. L. N. Kennett. Simulation and inversion of seismic wave propagation on continental scales based on a spectral-element method. *Journal of Numerical Analysis, Industrial and Applied Mathematics*, 4(1-2):11–22, June 2009.

[72] A. Fichtner, B. L. N. Kennett, H. Igel, and H.-P. Bunge. Full waveform tomography for upper-mantle structure in the Australasian region using adjoint methods. *Geophysical Journal International*, 179(3):1703–1725, 2009.

[73] H. P. Flath, L. C. Wilcox, V. Akçelik, J. Hill, B. van Bloemen Waanders, and O. Ghattas. Fast algorithms for Bayesian uncertainty quantification in large-scale linear inverse problems based on low-rank partial Hessian approximations. *SIAM Journal on Scientific Computing*, 33(1):407–432, 2011.

[74] W. Förstner and M. Boudewijn. A metric for covariance matrices. In *Quo vadis geodesia*, pages 113–128. University Stuttgart, 1999.

[75] W. Förstner and B. Moonen. A metric for covariance matrices. In *Geodesy-The Challenge of the 3rd Millennium*, pages 299–309. Springer, 2003.

[76] C. Fox, H. Haario, and J. A. Christen. Inverse problems. In *Bayesian Volume in Honour of Sir Adrian F.M. Smith*. Clarendon Press, 2012.

[77] S. Friedland and A. Torokhti. Generalized rank-constrained matrix approximations. *SIAM Journal on Matrix Analysis and Applications*, 29(2):656–659, 2007.

[78] D. Galbally, K. Fidkowski, K. Willcox, and O. Ghattas. Nonlinear model reduction for uncertainty quantification in large-scale inverse problems. *International Journal For Numerical Methods in Engineering*, 81:1581–1608, 2010.

[79] M. W. Gee, C. M. Siefert, J. J. Hu, R. S. Tuminaro, and M. G. Sala. ML 5.0 smoothed aggregation user's guide. Technical Report SAND2006-2649, Sandia National Laboratories, 2006.

[80] A. Gelman, J. B. Carlin, H. S. Stern, and D. B. Rubin. *Bayesian Data Analysis, Second Edition (Chapman & Hall/CRC Texts in Statistical Science)*. Chapman and Hall/CRC, 2 edition, July 2003.

[81] J. Geweke and H. Tanizaki. On Markov chain Monte Carlo methods for nonlinear and non-Gaussian state-space models. *Communications in Statistics-Simulation and Computation*, 28(4):867–894, 1999.

[82] J. Geweke and H. Tanizaki. Note on the Sampling Distribution for the Metropolis-Hastings Algorithm. *Communications in Statistics-Theory and Methods*, 32(4):775–789, 2003.

[83] R. G. Ghanem and P. D. Spanos. *Stochastic Finite Elements: A Spectral Approach*. Springer–Verlag, 1991.

[84] O. Ghattas, J. Martin, and G. Stadler. Bayesian UQ for global full waveform inversion. In preparation.

[85] W. R. Gilks, S. Richardson, and D. J. Spiegelhalter, editors. *Markov Chain Monte Carlo in practice*, volume 2. CRC press, 1996.

[86] M. Girolami and B. Calderhead. Riemann manifold Langevin and Hamiltonian Monte Carlo methods. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 73(2):123–214, 2011.

[87] J. W. Glen. The creep of polycrystalline ice. *Proceedings of the Royal Society of London, Series A, Mathematical and Physical Sciences*, 228(1175):519–538, 1955.

[88] D. N. Goldberg and O. V. Sergienko. Data assimilation using a hybrid ice flow model. *The Cryosphere*, 5:315–327, 2011.

[89] G. H. Golub and C. F. Van Loan. *Matrix Computations*. Johns Hopkins Studies in the Mathematical Sciences. Johns Hopkins University Press, Baltimore, MD, third edition, 1996.

[90] G. H. Golub and Q. Ye. An inverse free preconditioned krylov subspace method for symmetric generalized eigenvalue problems. *SIAM Journal on Scientific Computing*, 24(1):312–334, 2002.

[91] J. Goodman, K. K. Lin, and M. Morzfeld. Small-noise analysis and symmetrization of implicit monte carlo samplers. *Preprint*, 2014. arXiv preprint arXiv:1410.6151.

[92] R. Greve and H. Blatter. *Dynamics of ice sheets and glaciers*. Advances in Geophysical and Environmental Mechanics and Mathematics. Springer, 2009.

[93] M. D. Gunzburger. *Perspectives in Flow Control and Optimization*. SIAM, Philadelphia, 2003.

[94] H. Haario, M. Laine, M. Lehtinen, E. Saksman, and J. Tamminen. Markov chain Monte Carlo methods for high dimensional inversion in remote sensing. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 66:591–608, 2004.

[95] H. Haario, M. Laine, A. Miravete, and E. Saksman. DRAM: Efficient adaptive MCMC. *Statistics and Computing*, 16:339–354, 2006.

[96] H. Haario, E. Saksman, and J. Tamminen. An adaptive Metropolis algorithm. *Bernoulli*, 7(2):223–242, 2001.

[97] M. Hairer. Introduction to Stochastic PDEs. Lecture Notes, 2009.

[98] N. Halko, P.-G. Martinsson, and J. A. Tropp. Finding structure with randomness: Probabilistic algorithms for constructing approximate matrix decompositions. *SIAM Review*, 53(2):217–288, 2011.

[99] M. Hanke. *Conjugate gradient type methods for ill-posed problems*, volume 327. CRC Press, 1995.

[100] M. Hanke and P. C. Hansen. Regularization methods for large-scale problems. *Surv. Math. Ind*, 3(4):253–315, 1993.

[101] P. C. Hansen. Regularization, GSVD and truncated GSVD. *BIT Numerical Mathematics*, 29(3):491–504, 1989.

[102] P. C. Hansen. *Rank-Deficient and Discrete Ill-Posed Problems: Numerical Aspects of Linear Inversion*, volume 4. SIAM, 1998.

[103] K. M. Hanson and G. S. Cunningham. Posterior sampling with improved efficiency. In *Medical Imaging: Image Processing*, pages 371–382, 1998.

[104] W. K. Hastings. Monte Carlo sampling methods using Markov chains and their applications. *Biometrika*, 57(1):97–109, 1970.

[105] J. Heikkinen. Statistical inversion theory in X-ray tomography. Master's thesis, Lappeenranta University of Technology, Finland, 2008.

[106] E. Herbst. Gradient and Hessian-based MCMC for DSGE models. 2010. Unpublished manuscript.

[107] M. R. Hestenes. Pseudoinversus and conjugate gradients. *Communications of the ACM*, 18(1):40–43, 1975.

[108] M. R. Hestenes and E. Stiefel. *Methods of conjugate gradients for solving linear systems*, volume 49. National Bureau of Standards Washington, DC, 1952.

[109] D. Higdon, J. Gattiker, B. Williams, and M. Rightley. Computer model calibration using high-dimensional output. *Journal of the American Statistical Association*, 103(482):570–583, 2008.

[110] D. Higdon, H. Lee, and C. Holloman. Markov chain Monte Carlo-based approaches for inference in computationally intensive inverse problems. In J. M. Bernardo, M. J. Bayarri, J. O. Berger, A. P. Dawid, D. Heckerman, A. F. M. Smith, and M. West, editors, *Bayesian Statistics 7*, 2003.

[111] D. Higdon, C. S. Reese, J. D. Moulton, J. A. Vrugt, and C. Fox. *Handbook of Markov Chain Monte Carlo*, chapter Posterior Exploration for Computationally Intensive Forward Models. Chapman & Hall / CRC Press, 2010.

[112] M. Hinze, R. Pinnau, M. Ulbrich, and S. Ulbrich. *Optimization with PDE Constraints*. Springer, 2009.

[113] L. Homa, D. Calvetti, A. Hoover, and E. Somersalo. Bayesian preconditioned CGLS for source separation in MEG time series. *SIAM Journal on Scientific Computing*, 35(3):B778–B798, 2013.

[114] Y. Hua and W. Liu. Generalized Karhunen-Loève transform. *IEEE Signal Processing Letters*, 5(6):141–142, 1998.

[115] K. Hutter. *Theoretical Glaciology*. Mathematical Approaches to Geophysics. D. Reidel Publishing Company, 1983.

[116] W. James and C. Stein. Estimation with quadratic loss. In *Proceedings of the Fourth Berkeley Symposium on Mathematical Statistics and Probability*, volume 1, pages 361–379, 1961.

[117] J. Kaipio and E. Somersalo. *Statistical and Computational Inverse Problems*, volume 160 of *Applied Mathematical Sciences*. Springer-Verlag, New York, 2005.

[118] R. E. Kalman. A new approach to linear filtering and prediction problems. *Journal of Fluids Engineering*, 82(1):35–45, 1960.

[119] K. Karhunen. Über lineare methoden in der wahrscheinlichkeitsrechnung. *Ann. Acad. Sci. Fennicae. Ser. A. I. Math.-Phys*, 37:1–79, 1947.

[120] M. C. Kennedy and A. O'Hagan. Bayesian calibration of computer models. *Journal of the Royal Statistical Society. Series B (Statistical Methodology)*, 63(3):425–464, 2001.

[121] D. Komatitsch, S. Tsuboi, C. Ji, and J. Tromp. A 14.6 billion degrees of freedom, 5 teraflops, 2.5 terabyte earthquake simulation on the Earth Simulator. In *SC03: Proceedings of the International Conference for High Performance Computing, Networking, Storage, and Analysis*. ACM/IEEE, 2003.

[122] D. A. Kopriva. A conservative staggered-grid Chebyshev multidomain method for compressible flows. II. A semi-structured method. *Journal of Computational Physics*, 128(2):475–488, 1996.

[123] D. A. Kopriva, S. L. Woodruff, and M. Y. Hussaini. Computation of electromagnetic scattering with a non-conforming discontinuous spectral element method. *International Journal for Numerical Methods in Engineering*, 53(1):105–122, 2002.

[124] C. Lanczos. *An iteration method for the solution of the eigenvalue problem of linear differential and integral operators*. United States Governm. Press Office, 1950.

[125] M. Lassas, E. Saksman, and S. Siltanen. Discretization invariant Bayesian inversion and Besov space priors. *Inverse Problems and Imaging*, 3(1):87–122, 2009.

[126] K. J. H. Law. Proposals which speed up function-space MCMC. *Journal of Computational and Applied Mathematics*, 2013. In Press.

[127] L. LeCam. *Asymptotic Methods in Statistical Decision Theory*. Springer, 1986.

[128] E. Lehmann and G. Casella. *Theory of Point Estimation*. Springer-Verlag, 1998.

[129] R. B. Lehoucq, D. C. Sorensen, and C. Yang. *ARPACK users' guide: solution of large-scale eigenvalue problems with implicitly restarted Arnoldi methods*, volume 6. Siam, 1998.

[130] V. Lekić and B. Romanowicz. Inferring upper-mantle structure by full waveform tomography with the spectral element method. *Geophysical Journal International*, 185(2):799–831, 2011.

[131] A. Lewis. Derivatives of spectral functions. *Mathematics of Operations Research*, 21(3):576–588, 1996.

[132] F. Li, Q. Dai, W. Xu, and G. Er. Weighted subspace distance and its applications to object recognition and retrieval with image sets. *IEEE Signal Processing Letters*, 16(3):227–230, 2009.

[133] W. Li and O. A. Cirpka. Efficient geostatistical inverse methods for structured and unstructured grids. *Water Resources Research*, 42:W06402, 2006.

[134] E. Liberty, F. Woolfe, P. Martinsson, V. Rokhlin, and M. Tygert. Randomized algorithms for the low-rank approximation of matrices. *Proceedings of the National Academy of Sciences*, 104(51):20167, 2007.

[135] C. Lieberman, K. Fidkowski, K. Willcox, and B. van Bloemen Waanders. Hessian-based model reduction: large-scale inversion and prediction. *International Journal for Numerical Methods in Fluids*, 71(2):135–150, 2013.

[136] C. Lieberman, K. Willcox, and O. Ghattas. Parameter and state model reduction for large-scale statistical inverse problems. *SIAM Journal on Scientific Computing*, 32(5):2523–2542, 2010.

[137] F. Lindgren, H. Rue, and J. Lindström. An explicit link between gaussian fields and gaussian markov random fields: the stochastic partial differential equation approach. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 73(4):423–498, 2011.

[138] D. Lindley and A. Smith. Bayes estimates for the linear model. *Journal of the Royal Statistical Society. Series B*, pages 1–41, 1972.

[139] D. C. Liu and J. Nocedal. On the limited memory BFGS method for large scale optimization. *Mathematical programming*, 45(1-3):503–528, 1989.

[140] J. S. Liu. *Monte Carlo strategies in Scientific Computing*. Springer, New York, 2001.

[141] M. Loève. *Probability theory, Vol. II*, volume 46 of *Graduate Texts in Mathematics*. Springer-Verlag, Berlin, 4 edition, 1978.

[142] S. J. Marshall. Recent advances in understanding ice sheet dynamics. *Earth and Planetary Science Letters*, 240:191–204, 2005.

[143] J. Martin, L. C. Wilcox, C. Burstedde, and O. Ghattas. A stochastic Newton MCMC method for large-scale statistical inverse problems with application to seismic inversion. *SIAM Journal on Scientific Computing*, 34(3):A1460–A1487, 2012.

[144] Y. Marzouk and D. Xiu. A stochastic collocation approach to Bayesian inference in inverse problems. *Communications in Computational Physics*, 6(4):826–847, 2009.

[145] Y. M. Marzouk and H. N. Najm. Dimensionality reduction and polynomial chaos acceleration of Bayesian inference in inverse problems. *Journal of Computational Physics*, 228:1862–1902, 2009.

[146] Y. M. Marzouk, H. N. Najm, and L. A. Rahn. Stochastic spectral methods for efficient bayesian solution of inverse problems. *Journal of Computational Physics*, 224(2):560 – 586, 2007.

[147] J. C. Mattingly, N. Pillai, and A. M. Stuart. Diffusion limits of the random walk Metropolis algorithm in high dimensions. *Annals of Applied Probability*, 22:881–930, 2012.

[148] X. Meng, M. A. Saunders, and M. W. Mahoney. LSRN: A parallel iterative solver for strongly over-or underdetermined systems. *SIAM Journal on Scientific Computing*, 36(2):C95–C118, 2014.

[149] N. Metropolis, A. W. Rosenbluth, M. N. Rosenbluth, A. H. Teller, and E. Teller. Equation of state calculations by fast computing machines. *The Journal of Chemical Physics*, 21(6):1087–1092, 1953.

[150] E. H. Moore. On the reciprocal of the general algebraic matrix. *Bulletin of the American Mathematical Society*, 26:394–395.

[151] M. Morlighem, E. Rignot, H. Seroussi, E. Larour, H. Ben Dhia, and D. Aubry. Spatial patterns of basal drag inferred using control methods from a full-Stokes and simpler models for Pine Island Glacier, West Antarctica. *Geophysical Research Letters*, 37(14):L14502, 2010.

[152] R. M. Neal. *Handbook of Markov Chain Monte Carlo*, chapter MCMC using Hamiltonian dynamics. Chapman & Hall / CRC Press, 2010.

[153] N. Nguyen, G. Rozza, D. Huynh, and A. Patera. Reduced basis approximation and a posteriori error estimation for parametrized parabolic PDEs; Application to real-time Bayesian parameter estimation. In L. Biegler, G. Biros, O. Ghattas, M. Heinkenschloss, D. Keyes, B. Mallick, L. Tenorio, B. van Bloemen Waanders, and K. Willcox, editors, *Large Scale Inverse Problems and Quantification of Uncertainty*. John Wiley & Sons, 2011.

[154] J. Nocedal. Updating quasi-Newton matrices with limited storage. *Mathematics of computation*, 35(151):773–782, 1980.

[155] J. T. Oden, R. M. Moser, and O. Ghattas. Computer predictions with quantified uncertainty, Parts I & II. *SIAM News*, 43(9&10), 2010.

[156] D. S. Oliver, L. B. Cunha, and A. C. Reynolds. Markov chain Monte Carlo methods for conditioning a permeability field to pressure data. *Mathematical Geology*, 29(1):61–91, 1997.

[157] D. S. Oliver, A. C. Reynolds, and N. Liu. *Inverse theory for petroleum reservoir characterization and history matching*. Cambidge University Press, 2008.

[158] A. B. Owen. *Monte Carlo theory, methods and examples*. 2013.

[159] C. C. Paige and M. A. Saunders. Towards a generalized singular value decomposition. *SIAM Journal on Numerical Analysis*, 18(3):398–405, 1981.

[160] C. C. Paige and M. A. Saunders. Algorithm 583: LSQR: Sparse linear equations and least squares problems. *ACM Transactions on Mathematical Software (TOMS)*, 8(2):195–209, 1982.

[161] L. Pardo. *Statistical Inference Based on Divergence Measures*. CRC Press, 2005.

[162] W. S. B. Paterson. *The Physics of Glaciers*. Butterworth Heinemann, third edition, 1994.

[163] F. Pattyn, L. Perichon, A. Aschwanden, B. Breuer, B. de Smedt, O. Gagliardini, G. H. Gudmundsson, R. C. A. Hindmarsh, A. Hubbard, J. V. Johnson, T. Kleiner, Y. Konovalov, C. Martin, A. J. Payne, D. Pollard, S. Price, M. Ruckamp, F. Saito, O. Soucek, S. Sugiyama, and T. Zwinger. Benchmark experiments for higher-order and full-Stokes ice sheet models (ISMIP-HOM). *The Cryosphere*, 2(2):95–108, 2008.

[164] R. Penrose. A generalized inverse for matrices. In *Mathematical Proceedings of the Cambridge Philosophical Society*, volume 51, pages 406–413. Cambridge Univ Press, 1955.

[165] D. Peter, D. Komatitsch, Y. Luo, R. Martin, N. Le Goff, E. Casarotti, P. Le Loher, F. Magnoni, Q. Liu, C. Blitz, T. Nisson-Meyer, P. Basini, and J. Tromp. Forward and adjoint simulations of seismic wave propagation on fully unstructured hexahedral meshes. *Geophysical Journal International*, 186(2):721–739, 2011.

[166] N. Petra, J. Martin, G. Stadler, and O. Ghattas. A computational framework for infinite-dimensional Bayesian inverse problems: Part II. Stochastic Newton MCMC with application to ice sheet inverse problems. *SIAM Journal on Scientific Computing*, 36(4):A1525–A1555, 2014.

[167] N. Petra, H. Zhu, G. Stadler, T. J. R. Hughes, and O. Ghattas. An inexact Gauss-Newton method for inversion of basal sliding and rheology parameters in a nonlinear Stokes ice sheet model. *Journal of Glaciology*, 58(211):889–903, 2012.

[168] N. S. Pillai, A. M. Stuart, and A. H. Thiery. Optimal scaling and diffusion limits for the Langevin algorithm in high dimensions,. *Annals of Applied Probability*, 22:2320–2356, 2012.

[169] M. R. Pralong and G. H. Gudmundsson. Bayesian estimation of basal conditions on rutford ice stream, west antarctica, from surface data. *Journal of Glaciology*, 57(202):315–324, 2011.

[170] S. F. Price, A. J. Payne, I. M. Howat, and B. E. Smith. Committed sea-level rise for the next century from Greenland ice sheet dynamics during the past decade. *Proceedings of the National Academy of Sciences*, 108(22):8978, 2011.

[171] Y. Qi and T. P. Minka. Hessian-based Markov chain Monte-Carlo algorithms. In *First Cape Cod Workshop on Monte Carlo Methods*, Cape Cod, MA, USA, September 2002.

[172] M. J. Raymond and G. H. Gudmundsson. Estimating basal properties of ice streams from surface measurements: a non-linear Bayesian inverse approach applied to synthetic data. *The Cryosphere*, 3:265–278, 2009.

[173] J. Ritsema and J. Van Heijst. Constraints on the correlation of p-and s-wave velocity heterogeneity in the mantle from p, pp, ppp and pkpab traveltimes. *Geophysical Journal International*, 149(2):482–489, 2002.

[174] C. P. Robert and G. Casella. *Monte Carlo Statistical Methods (Springer Texts in Statistics)*. Springer-Verlag New York, Inc., Secaucus, NJ, USA, 2005.

[175] G. O. Roberts, A. Gelman, and W. R. Gilks. Weak convergence and optimal scaling of random walk Metropolis algorithms. *Annals of Applied Probability*, 7:110–120, 1997.

[176] G. O. Roberts and J. S. Rosenthal. Optimal scaling of discrete approximations to Langevin diffusions. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 60:255–268, 1998.

[177] G. O. Roberts and J. S. Rosenthal. Optimal scaling for various metropolis-hastings algorithms. *Statistical Science*, 16(4):pp. 351–367, 2001.

[178] G. O. Roberts and J. S. Rosenthal. Coupling and ergodicity of adaptive Markov chain Monte Carlo algorithms. *Journal of Applied Probability*, 44(2):458–475, 2007.

[179] G. O. Roberts and J. S. Rosenthal. Examples of adaptive MCMC. *Journal of Computational and Graphical Statistics*, 18(2):349–367, 2009.

[180] G. O. Roberts and R. L. Tweedie. Exponential convergence of Langevin distributions and their discrete approximations. *Bernoulli*, 2(4):341–363, 1996.

[181] H. Rue. Fast sampling of gaussian markov random fields. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 63(2):325–338, 2001.

[182] H. Rue and L. Held. *Gaussian Markov random fields: theory and applications*, volume 104. Chapman & Hall, 2005.

[183] Y. Saad. *Iterative methods for sparse linear systems*. Society for Industrial and Applied Mathematics, Philadelphia, PA, second edition, 2003.

[184] A. H. Sameh and J. A. Wisniewski. A trace minimization algorithm for the generalized eigenvalue problem. *SIAM Journal on Numerical Analysis*, 19(6):1243–1259, 1982.

[185] A. Solonen, H. Haario, J. Hakkarainen, H. Auvinen, I. Amour, and T. Kauranne. Variational ensemble Kalman filtering using limited memory BFGS. *Electronic Transactions on Numerical Analysis*, 39:271–285, 2012.

[186] A. Spantini, A. Solonen, T. Cui, J. Martin, L. Tenorio, and Y. Marzouk. Optimal low-rank approximations of Bayesian linear inverse problems. *Submitted*, 2014. arXiv preprint arXiv:1407.3463.

[187] G. Stewart. The efficient generation of random orthogonal matrices with an application to condition estimators. *SIAM Journal on Numerical Analysis*, 17:403–409, 1980.

[188] G. Strang and G. J. Fix. *An Analysis of the Finite Element Method*. Wellesley-Cambridge Press, Wellesley, MA, 1988.

[189] A. M. Stuart. Inverse problems: A Bayesian perspective. *Acta Numerica*, 19:451–559, 2010.

[190] A. M. Stuart, J. Voss, and P. Wiberg. Conditional path sampling of SDEs and the Langevin MCMC method. *Communications in Mathematical Sciences*, 2(4):685–697, 2004.

[191] C. Tape, Q. Liu, A. Maggi, and J. Tromp. Adjoint tomography of the southern California crust. *Science*, 325:988–992, 2009.

[192] A. Tarantola. *Inverse Problem Theory and Methods for Model Parameter Estimation*. SIAM, Philadelphia, PA, 2005.

[193] L. Tierney. Markov chains for exploring posterior distributions. *The Annals of Statistics*, 22(1):1701–1762, 1994.

[194] A. Tikhonov. Solution of incorrectly formulated problems and the regularization method. In *Soviet Math. Dokl.*, volume 5, pages 1035–1038, 1963.

[195] F. Tröltzsch. *Optimal Control of Partial Differential Equations: Theory, Methods and Applications*, volume 112 of *Graduate Studies in Mathematics*. American Mathematical Society, 2010.

[196] H. J. van Heijst, J. Ritsema, and J. H. Woodhouse. Global P and S velocity structure derived from normal mode splitting, surface wave dispersion and body wave travel time data. In *Eos Trans.*, page S221. AGU, 1999.

[197] C. F. Van Loan. Generalizing the singular value decomposition. *SIAM Journal on Numerical Analysis*, 13(1):76–83, 1976.

[198] C. R. Vogel. *Computational Methods for Inverse Problems*. Frontiers in Applied Mathematics. Society for Industrial and Applied Mathematics (SIAM), Philadelphia, PA, 2002.

[199] J. Voss. The effect of finite element discretisation on the stationary distribution of SPDEs. *Communications in Mathematical Sciences*, 10(4):1143–1159, 2012.

[200] J. Wang and N. Zabaras. Using Bayesian statistics in the estimation of heat source in radiation. *International Journal of Heat and Mass Transfer*, 48(1):15 – 29, 2005.

[201] L. C. Wilcox, G. Stadler, C. Burstedde, and O. Ghattas. A high-order discontinuous Galerkin method for wave propagation through coupled

elastic-acoustic media. *Journal of Computational Physics*, 229(24):9373–9396, 2010.

[202] A. T. A. Wood and G. Chan. Simulation of stationary Gaussian processes in $[0, 1]$d. *Journal of computational and graphical statistics*, 3(4):409–432, 1994.

[203] S. J. Wright and J. Nocedal. *Numerical Optimization*, volume 2. Springer New York, 1999.

[204] Y. Yue and P. L. Speckman. Nonstationary spatial Gaussian Markov random fields. *Journal of Computational and Graphical Statistics*, 19(1), 2010.

[205] N. Zabaras and B. Ganapathysubramanian. A scalable framework for the solution of stochastic inverse problems using a sparse grid collocation approach. *Journal of Computational Physics*, 227(9):4697 – 4735, 2008.

[206] H. Zhu, E. Bozdağ, D. Peter, and J. Tromp. Structure of the European upper mantle revealed by adjoint tomography. *Nature Geoscience*, 5:493–498, 2012.